Learning to Reformulate Long Queries for Clinical Decision Support*

Luca Soldaini
Information Retrieval Lab, Georgetown University, Washington, DC, USA
luca@ir.cs.georgetown.edu

Andrew Yates
Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbruecken, Germany
ayates@mpi-inf.mpg.de

Nazli Goharian
Information Retrieval Lab, Georgetown University, Washington, DC, USA
nazli@ir.cs.georgetown.edu

Abstract

The large volume of biomedical literature poses a serious problem for medical professionals, who are often struggling to keep current with it. At the same time, many health providers consider knowledge of the latest literature in their field a key component for successful clinical practice. In this work, we introduce two systems designed to help retrieving medical literature. Both receive a long, discursive clinical note as input query, and return highly relevant literature that could be used in support of clinical practice. The first system is an improved version of a method previously proposed by the authors; it combines pseudo relevance feedback and a domain specific term filter to reformulate the query. The second is an approach that uses a deep neural network to reformulate a clinical note. Both approaches were evaluated on the 2014 and 2015 TREC CDS datasets; in our tests, they outperform the previously proposed method by up to 28% in inferred NDCG; furthermore, they are competitive with the state of the art, achieving up to 8% improvement in inferred NDCG.

*This is a preprint of an article accepted for publication in Journal of the Association for Information Science and Technology

A 46-year-old woman presents with a 9 month history of weight loss (20 lb), sweating, insomnia and diarrhea. She reports to have been eating more than normal and that her heart sometimes races for no reason. On physical examination her hands are warm and sweaty, her pulse is irregular at 110bpm and there is hyperreflexia and mild exophthalmia.

Figure 1: An example of a query in our dataset (#6, TREC 2015).

Introduction

Keeping up-to-date with current literature is often a challenging task for medical professionals, especially for those who spend the majority of their time practicing. Surveys showed being knowledgeable about the latest research findings is a key component of good clinical practice (Tenopir et al., 2007; Naveh et al., 2015); yet, many physicians report that they do not read as much as they feel they should (Burke et al., 2004). Nowadays, PubMed's search tool[1] is widely used to search medical literature; however, it only supports boolean queries, which limits its use cases. Understandably, interest in systems designed to retrieve medical literature has increased in the past two decades. For example, many test collections have been introduced to advance the state of the art in medical search systems: OHSUMED (Hersh et al., 1994) focused on retrieving literature for short, keyword-heavy queries; TREC Genomics (Hersh and Voorhees, 2009) tackled search in support of genomics research; ImageCLEFmed (Kalpathy-Cramer et al., 2015) studied multimodal retrieval for clinical practice; MedTrack (Voorhees and Tong, 2011; Voorhees and Hersh, 2012) was concerted with improving retrieval of clinical notes.

In 2014, the Clinical Decision Support shared task was introduced at the Text REtrieval Conference[2] (TREC) (Roberts et al., 2016a). The goal of the task is to study the problem of literature retrieval in support of clinical practice; that is, task participants were invited to create systems that, given a clinical note describing the conditions of a patient, would retrieve highly relevant medical literature from the Open Access Subset of PubMed that could help making a diagnosis and/or determine a treatment. This would reduce the time a medical professional needs to spend formulating boolean queries to retrieve the most up-to-date studies about conditions and treatments pertinent to their patients. Compared to other biomedical search tasks, Clinical Decision Support search (CDS search) is characterized by long, discursive queries (on average, each clinical note has more than 80 terms) written by medical experts. Each query contains biographical information, patient history, and current medications; an example is shown in Figure 1.

The shared task captured the interest of many research teams, and ran again the following year (Roberts et al., 2016b). Among the many approaches proposed, automatic query expansion techniques were found to be very effective for the task [e.g., (Choi and Choi, 2014; Mourao et al., 2014; Balaneshin Kordan et al., 2015).] Some expansion techniques relied on medical ontologies; others were based on Pseudo Relevance Feedback (PRF), a method in information retrieval to expand queries by selecting m number of good terms from k number of top ranked documents.

[1] `http://www.ncbi.nlm.nih.gov/pubmed/advanced`

[2] `http://trec.nist.gov/`

In this work, we introduce two systems designed for CDS search. Both systems reformulate long, discursive queries by adding relevant terms to the information need expressed in the query. The first method — an improved version of (Soldaini et al., 2014) and (Soldaini et al., 2015) — expands the query by using pseudo relevance feedback; then it prunes the list of expansion candidates by removing those that are not medically related. The second method is a supervised approach to query expansion; it uses a multi-layer neural network to predict, given a list of possible candidate terms, which terms to add to the original query to improve document retrieval.

Finally, we study the effect of query reduction when combined with query expansion; several methods — one taking advantage of term distribution on an external collection, the other leveraging syntactic analysis — are compared.

In summary, our contributions are as follows:

- We introduce two methods for query expansion for CDS search.
- We compared the proposed methods with the current state of the art.
- We study the impact of query reduction when combined with query expansion in CDS search.

Related Works

Search in the health domain has been a topic of interest for more than two decades. Over the years, many systems that rely on query reformulation have been proposed to improve retrieval in this domain. In this section, we present an overview of query reformulation techniques applied to various ad-hoc search tasks in the health domain.

OHSUMED Early on, Hersh et al. (1994) introduced the OHSUMED test collection, a dataset comprised of 106 queries and 348,566 documents. The collection was created to promote the comparison of search systems for biomedical search. Documents consisted of a subset of articles from MEDLINE[3], a large-scale repository of biomedical journal citations and abstracts; queries were generated by 22 medical professionals (11 librarians and 11 physicians) who were already familiar with MEDLINE. Queries in the dataset are, on average, 14 terms long, which is much shorter than the queries considered in this manuscript (80 terms). After its introduction, the OHSUMED collection has been extensively used to evaluate classification [e.g., (Genkin et al., 2007; Han and Karypis, 2000; Xu and Li, 2007)], learning to rank [e.g. (Cao et al., 2006; Duh and Kirchhoff, 2008; Liu et al., 2007)], and query reformulation (Abdou and Savoy, 2008; Dong et al., 2011; Haveliwala, 2002; Hersh et al., 2000; Jalali and Borujerdi, 2011; Liu and Chu, 2007; Srinivasan, 1996; Thesprasith and Jaruskulchai, 2014). Works in the latter group are the most similar to our systems; they can be further partitioned based on the approach used: ontology-based reformulation, Pseudo Relevance Feedback (PRF), and a combination of the two. Early on, Srinivasan (1996) introduced SMART, a retrieval system that uses the MeSH ontology[4]—a controlled vocabulary used by the US National Library of Medicine to tag and index articles in PubMed—to expand a query. Two experiments

[3] `https://www.nlm.nih.gov/bsd/pmresources.html`

[4] `https://www.ncbi.nlm.nih.gov/mesh`

were carried out: in the first, MeSH terms were used to expand the query; in the second, they were used as search terms to retrieve documents to perform PRF expansion. A two step approach—MeSH expansion first, then PRF expansion—was also tested; overall, the system improved up to 17% over a Vector Space Model (VSM) (Salton et al., 1975) baseline. Hersh et al. (2000) expanded queries with terms manually selected from the UMLS Metathesaurus[5] relationships to enhance retrieval performance; experimental results showed that thesaurus based query expansion did not always improve search efficiency. More recently, Liu and Chu (2007) also used UMLS to perform query expansion; their system automatically expands the query using scenario-specific terms (where a scenario could be "make a diagnosis" or "finding a treatment"). The system works in two steps: first, UMLS terms that occur frequently with terms in the query are identified. Second, scenario-specific terms are identified using the neighbors of concepts in the query in the UMLS graph whose semantic type matches (e.g. semantic type "Disease or Syndrome" for query scenario "make a diagnosis"). Abdou and Savoy (2008) introduced a variant of the Rocchio query expansion formula (Rocchio, 1971) for search in MEDLINE; their system improved up to 13.5% over SMART. Jalali and Borujerdi (2011) proposed a method that incorporates medical concepts in the PRF process. In detail, MeSH terms are used in conjunction with query terms to rank MEDLINE documents. Dong et al. (2011) adapted PageRank to perform query expansion using the UMLS ontology. Specifically, terms in UMLS are used as nodes for the PageRank; relationships between concepts are used to determine popularity. A variant of PageRank (Haveliwala, 2002) that takes into account the popularity of UMLS terms in the OHSUMED collection was used to bias the concepts network. At query time, terms in the query are mapped to UMLS concepts; highly ranked concepts related to query concepts are then used for query expansion. Finally, Thesprasith and Jaruskulchai (2014) introduced RABAM-PRF, a variant of pseudo relevance feedback that ranks MeSH terms found in the top documents and uses them for query expansion.

Overall, both statistical (i.e., PRF) and thesaurus-based query expansion techniques have been proven effective on the OHSUMED dataset; however, while some have found the former to outperform the latter (Jalali and Borujerdi, 2011), others have reached the opposite conclusion (Dong et al., 2011; Liu and Chu, 2007), or determined that a combination of the two is the most effective strategy (Srinivasan, 1996).

TREC Genomics Track Between 2003 and 2007, the Genomics Track at TREC (Hersh and Voorhees, 2009) promoted the study of new approaches for searching biology literature. TREC Genomics started as an ad-hoc retrieval task, later including summarization, text categorization, and question answering tasks. Approaches proposed for this task do not directly compare with our system because of differences in the document collection (biology instead of medical domain) and in the queries of the ad-hoc task (9 terms long on average vs 82 in our dataset; futhermore, the queries are keyword-heavy). Some have studied how to adapt retrieval model for this task [e.g., (Urbain et al., 2009)], while others focused on query expansion techniques [e.g., (Lu et al., 2009; Matos et al., 2010; Stokes et al., 2009)]. Hersh and Voorhees (2009) noted that, among the groups who participated in ad-hoc retrieval task, those who used domain-specific query expansion (e.g.,

[5]The UMLS Metathesaurus is a large collection of biomedical and health-related concepts, their synonymous names, and their relationships; `https://www.nlm.nih.gov/research/umls/`

synonym based expansion) achieved the best performance (Büttcher et al., 2004), while pseudo relevance feedback methods were found beneficial in re-weighting terms in the query (Zheng et al., 2005). Subsequent works have confirmed these findings; for example, Stokes et al. (2009) noted that "query expansion has a positive effect on genomic retrieval performance ... [but] expansion terms should be gleaned for manually-derived domain specific resources." Similarly, Lu et al. (2009) and Matos et al. (2010) proposed concepts-based query expansions systems.

TREC MedTrack Track Retrieval of medical records has been evaluated as part as the 2011 and 2012 MedTrack at TREC (Voorhees and Tong, 2011; Voorhees and Hersh, 2012). In detail, participants were asked to retrieve clinical records of patients matching a criteria expressed in a query. The track only ran for two years, due to a "lack [of] a suitable collection of health records to serve as the basis of a test collection" (Voorhees, 2013). Furthermore, the collection is not available to those that have not participated to MedTrack due to privacy concerns. Nevertheless, the collection still attracts the interest of many researchers. Some have investigated how to exploit semantic relationship between terms in the query and terms in the documents. For example, Choi et al. (2014) introduced a ranking method that uses a semantic concept-enriched dependence model: documents containing medical concepts that appear in close proximity in the query receive a high similarity score. Recently, Koopman et al. (2016) have proposed an inference model to address the semantic gap between queries and medical records. Their approach uses a combination of statistical features and domain knowledge to define a graph on which the inference mechanism is applied. Then, given a query, each document is scored based on the amount of evidence supporting the relationship between query and document. Other have investigated the use of query expansion to improve retrieval; Limsopatham et al. (2013) uses a combination of medical concepts extracted from the top retrieved documents and concept relationship obtained from ontologies and external collections to expand the query. Similarly, Zhu et al. (2014) explored the use of four auxiliary collections of clinical records, medical literature, and general domain web pages to build a mixture of relevance model for query expansion. They observed that a combination of all resources lead to the largest improvement over a query likelihood baseline. Moreover, they noted that the largest improvements were obtained on low-performing queries. Overall, works in this area suggest that exploiting external collections is an effective approach for clinical notes retrieval, at least in the context of the MedTrack shared task.

ImageCLEF Med Between 2009 and 2013, one of the tasks of ImageCLEFmed (de Herrera et al., 2013; Kalpathy-Cramer et al., 2011; Müller et al., 2009, 2010, 2012) asked participants to retrieve, for a given clinical note, papers from PubMed that described similar cases. This task presents significant differences with the problem studied in this paper. First, clinical notes are, on average, shorter than the clinical notes in our dataset (43 vs 82 terms on average); in fact, they are much similar to the summaries provided alongise clinical notes in the CDS TREC datasets. Second, images were provided alongside each query, as the task was conceived as a multimodal retrieval task. Finally, the document collection was restricted to just case report literature (that is, just publications reporting a medical case) rather than using the full open access subset of PubMed. Many teams experimented with query expansion techniques for the textual component

of the retrieval systems: Choi and Choi (2013) used the top documents retrieved from an auxiliary collection of medical documents (MEDLINE) to perform pseudo relevance feedback; Kitanovski et al. (2013) combined both term and medical concept pseudo relevance feedback, achieving up to 25% improvement over the non-expanded queries; Mourao et al. (2013) improved their retrieval method by expanding the query with MeSH terms; Simpson et al. (2013) also used medical concepts to expand the textual queries. Overall, organizers noted query expansion techniques that exploited medical thesauri and databases were the most effective (Kalpathy-Cramer et al., 2015).

CLEF eHealth More recently, a task concerned with improving systems designed to help laypeople seeking health information online was introduced in the ShARe/CLEF eHealth Evaluation Lab (Goeuriot et al., 2013, 2014). Instead of biomedical literature, participating systems were asked to retrieve relevant documents from a set of approved websites by the Health On the Net (HON) Foundation[6]—an organization that certifies those health-related websites that meet specific reliability standards— and other hand-picked trusted resources. The task is modeled from the point of view of a lay person with no medical experience, who has different information needs of healthcare professional; in contrast, in this manuscript we aim at retrieving medical literature for medical experts; thus, the task introduced in CLEF eHeath does not directly map to the problem studied in this manuscript. Nevertheless, similar query expansion techniques to those mentioned in the previous task were proposed: pseudo relevance feedback, Mixture of Relevance Models (Diaz and Metzler, 2006), and expansion using medical ontologies and databases (e.g., UMLS). The 2013 task was proven to be very challenging, as only one team (Zhu et al., 2013) was able to outperform the simple yet competitive BM25 baseline with pseudo relevant feedback introduced by the task organizers (P@10 = 0.4860). In detail, Zhu et al. (2013) explored the use of the MeSH ontology for query expansion, as well as an Mixture of Relevance Models approach. Their best run uses a Markov Random Field model (Metzler and Croft, 2005) for retrieval and Mixture of Relevance Models query expansion using four collections (CLEF eHealth, TREC Medical, TREC Genomics, and a set of clinical notes from MayoClinic), achieving a 9% improvement over the baseline, although the difference is not statistically significant (Mann-Whitney U test, $p \geq 0.05$). Interestingly, teams who took advantage of medical ontologies did not perform better than the baseline [e.g., (Choi and Choi, 2013; Bedrick and Sheikshabbafghi, 2013)]. Conversely, on the 2014 dataset, systems who took advantage of medical resources outperformed the baseline (P@10 = 0.68) significantly, perhaps due to the fact that the dataset from the previous year could be used for training and tuning. Shen et al. (2014) considered a concept-based similarity model; MetaMap (Aronson and Lang, 2010) was used to extract medical concepts from the queries and documents; furthermore, the authors experimented with using concept-based pseudo relevance feedback. Their best approach also resulted in a 11% improvement over the baseline. Oh and Jung (2014) used a combination of rule-based expansion of medical abbreviations, expansion through terms in the clinical notes, and pseudo relevance feedback. Their system achieved a 8% improvement over the baseline. Overall, we note how both thesauri-based query expansion and pseudo relevance feedback techniques have been proved successful for this task; however, for both years, proposed methods achieved limited improvements over a strong baseline.

[6] `http://www.healthonnet.org`

TREC CDS track In recent years, the Clinical Decision Support (CDS) track was introduced at TREC (Roberts et al., 2016a,b) with the goal of promoting the study of systems to "provide relevant articles to clinicians to improve their decision-making in diagnosing, treating, and testing patients." Compared to the datasets mentioned above, the novelty of this track is in the fact that — instead of queries — clinical narratives are used to describe the information need. In other words, systems receive a clinical note consisting of several sentences as input rather than a short, keyword-heavy query. Query reformulation techniques were extensively studied by most participating teams in CDS TREC 2014 and 2015 (Balaneshin Kordan et al., 2015; Choi and Choi, 2014; Cohan et al., 2014; Jiang et al., 2016; McNamee, 2015; Mourao et al., 2014; Sankhavara et al., 2014; Sierek and Hanbury, 2015; Soldaini et al., 2014, 2015; Xu et al., 2014); some employed medical lexica and thesauri, while others used pseudo relevance feedback techniques. Furthermore, many teams evaluated both.

Among those who took advantage of medical thesauri, Mourao et al. (2014) used MeSH terms to expand the query; the modified query was then used to retrieve and rank documents using multiple scoring functions [BM25L, BM25+ (Lv and Zhai, 2011), tf-idf, language model with Dirichlet smoothing (Zhai and Lafferty, 2001)]. Finally, the rank of retrieved documents was determined by combining the ranks given by each scoring function using the Reciprocal Rank Fusion algorithm (Cormack et al., 2009). Balaneshin Kordan et al. (2015) used Markov Random Field Parameterized Query Expansion, a mixture model that weights terms based on whether they appeared in the query, in top retrieved documents, or in the UMLS ontology.

Many explored the use of pseudo relevance feedback for CDS. For example, Choi and Choi (2014) used titles, abstracts, and MeSH terms from the MEDLINE collection to obtain expansion terms for each query. Documents retrieved by the expanded query were then re-ranked using three classifiers trained to identify papers that matched the scenario. Xu et al. (2014) and McNamee (2015) combined HAIRCUT (McNamee and Mayfield, 2004), a character n-grams search engine, with pseudo relevance feedback. Their system achieved a 25% increase in inferred Normalized Discounted Cumulative Gain (infNDCG) (Yilmaz et al., 2008) when PRF is used over their non-expanded baseline. Sankhavara et al. (2014) compared pseudo relevance feedback with manual relevance feedback. The Terrier search engine[7] was used to perform PRF. Surprisingly, the two techniques achieved similar results. Oh and Jung (2015) proposed a method that employs external collections to generate candidate terms to add to the queries. Documents retrieved from external collections are clustered; terms from each cluster are then employed to expand the query. The proposed method was tested on three collections: TREC CDS, OHSUMED, and CLEF eHealth; however, it achieved statistically significant improvement over a language model baseline in the first two cases (+10.32% and +12.33% respectively).

Some researchers have also investigated other means of performing query expansion. For example, Jiang et al. (2016) studied the topology of the network of publications in the Open Access Subset of PubMed. In the proposed graph, each node represent a document; documents share an edge if one or more medical concepts co-occur in both documents. In their analysis, they observed that relevant papers for a query tend to cluster together; therefore, they proposed a re-ranking algorithm that promotes documents based on which clusters they belong to.

[7] `http://terrier.org/`

Finally, the authors of this manuscript have also explored the use of pseudo feedback techniques. In (Cohan et al., 2014; Soldaini et al., 2014, 2015), we used a feedback technique similar to (Abdou and Savoy, 2008) to obtain terms suitable for query expansion. Terms were then filtered based on their likelihood of appearing in health-related Wikipedia pages. As previously mentioned, we introduce an improved version of this algorithm in this manuscript, as well as a supervised approach to query expansion.

As evidence by the large body of research presented in this section, several common approaches exist among the many systems proposed in the last two decades. For query expansion, most systems have relied on statistical query expansion (e.g, PRF), query expansion through domain specific ontologies, and query expansion through auxiliary collections. In particular, statistical approaches seem to be more effective in case of short queries [e.g., (Jalali and Borujerdi, 2011)], as it is the case for OHSUMED, while techniques that take advantage of domain specific resources are more effective in the case of longer and more complex queries [e.g., (Balaneshin Kordan et al., 2015; Choi and Choi, 2013; Lu et al., 2009; Stokes et al., 2009; Simpson et al., 2013; Soldaini et al., 2015; Zhu et al., 2014)]. Interestingly, the use of auxiliary collections has been exploited for both clinical notes and medical and non-medical literature; however, further studies are needed to assess which characteristics make an auxiliary collection suitable for query expansion: for example, Zhu et al. (2014) found that general domain collections improve the performance of the overall systems, while Oh and Jung (2015) concluded that non-medical documents have a negative impact on some tasks.

In recent years, some domain agnostic query reduction have been proposed. For example, Kumaran and Carvalho (2009) used a learning to rank approach to find the best sub-query using a series of clarity predictors and similarity measures as features. The proposed method was found not to perform well in case of long, discursive queries such as case reports (Soldaini et al., 2015). Bendersky and Croft (2008) used a supervised method for identifying key concepts in long queries; in a subsequent work, they assigned different weights to concepts extracted from the query (Bendersky et al., 2010). The framework introduced in the latter work inspired the system introduced by (Balaneshin Kordan et al., 2015) for CDS search. These examples suggest that domain agnostic methods for query reduction need to be revisited to be effective for the task studied in this manuscript.

Methodology

As documented in the previous section, researchers have shown that query reformulation techniques are very effective at improving retrieval performance of CDS search systems.

Informed by such findings, we propose a three-stage approach to reformulate long, discursive queries. The first stage takes advantage of the PRF method introduced in (Soldaini et al., 2015) to generate term candidates. In the second stage, a subset of candidate terms are selected for query expansion. Two candidate selection methods are compared: the first is an improved version of Health Terms Pseudo Relevance Feedback (HTPRF) (Soldaini et al., 2015); the second is a supervised approach. Finally, in the third stage, the query is expanded using the terms selected

in the previous step; furthermore, we also experimented with statistical and syntactical query reduction methods to remove terms from the query that could cause query drift.

Candidates Generation

Candidate terms for query expansion are generated using the pseudo relevance feedback method introduced in (Soldaini et al., 2015). For each query, the algorithm assigns a score s_j to each term t_j appearing in the k highest ranked documents.

In detail, the method works as follows: given a query Q and a document collection $\mathcal{D}$, it firstly retrieves and tokenizes k documents $\{d_1, \ldots, d_k\}$ from document collection $\mathcal{D}$; then, it builds the root set of query Q; that is, it generates the set $\mathcal{P}_Q$ of all terms appearing in any of the documents $\{d_i, \ldots, d_k\}$. Each term $t_j \in \mathcal{P}_Q$ is associated with a score s_j defined as follows:

$$\begin{aligned} s_j &= \log_{10}(10 + w_j) \\ w_j &= \alpha \cdot tf(t_j, Q) + \frac{\beta}{k} \sum_{i=1}^{k} tf(t_j, D_i) \cdot idf(t_j, \mathcal{D}) \end{aligned} \tag{1}$$

where $tf(t_j, Q)$ is the term frequency of term t_j in Q, $tf(t_j, D_i)$ is the term frequency of term t_j in document d_i, and $idf(t_j, \mathcal{D})$ is the inverse document frequency of the j-th term in the collection $\mathcal{D}$, as defined in (Grossman and Frieder, 2012, ch. 2). α and β are smoothing factors; the value of w_j is increased by ten before calculating s_j to ensure that all scores are greater or equal to one.

In our implementation, the top 500 candidate terms ranked by s_j are considered for query expansion. This choice is due to efficiency reason and does not impact the performance of the system, as the final number of expansion terms is, in all experiments, an order of magnitude smaller.

In our experiments, we found the scoring method shown in Equation 1 is quite stable with respect to the choice of parameters α and β: increasing or decreasing either of the two parameters by up to an order of magnitude causes little variation in the performance of the system. Therefore we set $\alpha = 2.0$ and $\beta = 0.75$ as suggested in (Soldaini et al., 2014, 2015). On the other hand, the number of top documents k does affect the retrieval performance of the algorithm; therefore, we will discuss the tuning of this parameter in the results section.

HTPRF Candidate Selection

Introduced in (Soldaini et al., 2015), HTPRF takes into account the likelihood of each candidate term of being health-related to determine whether to include it in the reformulated query. This likelihood is estimated using the set of health-related Wikipedia pages.

Let $\mathcal{W} = \{P_i\}_{i=1}^{i=|\mathcal{W}|}$ be the set of all pages in English Wikipedia (special pages, such as category or disambiguation pages, are not included), $\mathcal{W}_H$ the set of all health-related pages. Then, for each candidate term $t_j \in \mathcal{T}$, we estimate its odds ratio of being health related as follows:

$$\mathrm{OR}(t_j) = \frac{\Pr\{t_j \in P_i \wedge P_i \in \mathcal{W}_H\}}{\Pr\{t_j \in P_i \wedge P_i \in \mathcal{W}\}} \tag{2}$$

Figure 2: The Wikipedia entry for "Gastroesophageal reflux disease". The information box — highlighted in orange — contains several medically-related identification codes (ICD-10, ICD-9, OMIM, DiseaseDB, MedlinePlus, eMedicine, MeSH); thus, the page was identified as health-related.

The two probabilities are estimated using Maximum Likelihood Estimation (MLE); that is, they are calculated by dividing the number of documents with term t_j by the total number of documents. A candidate term t_j is kept if and only if $\text{OR}(t_j) \geq \delta$, where δ is a tuning parameter of our system. Of the remaining candidate terms, the top m ranked by s_j are considered for query expansion. As with k, the value of δ and m influence the performance of the retrieval algorithm; we analyze the effect of different values for δ and m in the results section.

A Wikipedia crawl from May 5, 2016 (5,116,922 pages) was used to compute above probabilities. We considered any page containing an information box with one of the following medically-related fields as a health-related page: MedlinePlus, DiseasesDB, eMedicine, MeSH, or OMIM (for a total of 22,943 pages). An example of the heuristic used to identify an health-related page is shown in Figure 2.

Deep Neural Network (DNN) Supervised Candidate Selection

We also approached query expansion as a supervised learning task where the goal is to predict which candidate terms should be used to expand the query. After candidate terms have been generated in step 1, we train a deep neural network to predict each candidate term's Weight Relevance Ratio (WRR), a value that represent the importance of a term in relevant documents. We used three groups of features to train our supervised model: word embedding representations of the query and terms, statistical features over multiple auxiliary collections, and other syntactical and semantic features. Word embeddings, a means of representing terms from a vocabulary into a dense, low-

dimensionality space, were obtained using the `word2vec` model (Mikolov et al., 2013). We detail the statistical features over external collections, as well as syntactical and semantic features in a later section; we will refer to them as "candidate features" (opposed to "candidate word embedding") throughout the rest of the manuscript.

The WRR of a candidate is defined as the ratio of its probability of appearing in a relevant document over its probability of appearing in the entire collection, weighted by its own frequency in the relevant documents. Similarly to (Mengle and Goharian, 2009), we found odds ratio to be a reliable indicator of importance in the relevant category. We scale the odds ratio of each term to prevent extremely rare terms from having a very high WRR score, as we empirically noticed that such terms are often spelling errors or non-relevant terms. Formally, given a term t, a collection $\mathcal{D} = \{P_i\}_{i=1}^{i=|\mathcal{D}|}$ of documents, and the set $\mathcal{R}_Q$ of relevant documents for query Q, $\mathcal{R}_Q \subset \mathcal{D}$, we defined WRR as follows:

$$\text{WRR}(t) = \log_{10}\left(cf(t, \mathcal{R}_Q) + 1\right) \cdot \frac{\Pr\{t \in P_i \wedge P_i \in \mathcal{R}_Q\}}{\Pr\{t \in P_i \wedge P_i \in \mathcal{D}\}} \tag{3}$$

We note that we scale the collection frequency $cf(t, \mathcal{R}_Q)$ of term t in set of relevant documents $\mathcal{R}_Q$ by taking its log to prevent very frequent terms from having a high WRR. The two probabilities are estimated using MLE, i.e. by dividing the number of documents with term t by the total number of documents. We predict WRR using a regression with mean squared error (MSE) as the loss function.

Our neural network consists of two components: a component that learns query and term representations in order to compute the similarity between them, and a component that predicts the candidate term's WRR based on the terms similarity with the query and the candidate term's features, which are described in the features section. This design is modeled after the neural network proposed by (Severyn and Moschitti, 2015), which learns query and document representations in order to rerank pairs of short documents (i.e., pairs of sentences and pairs of tweets). Our model primarily differs in that we use a single dense layer to learn term representations, whereas Severyn and Moschitti use a convolutional network to learn representations of the sequences of terms in two short documents. This change is due to that fact that, unlike their work, our system predicts the score of a single candidate rather than a passage.

In the next section we describe the neural network's query-term similarity component in detail; we describe the model's parameters later in the experimental setup section. The purpose of the second component of our neural network model is to combine the query-term similarity with additional features in order to make a WRR prediction; it consists of two layers: (*i*) a dense (i.e., fully connected) layer that takes the query-term similarity and candidate features as input (shown as *query-term similarity* and *features* in Figure 3) and filters them with a ReLU activation function (Nair and Hinton, 2010), and (*ii*) a dense layer that takes the previous dense layer's output as input and predicts the term's WRR (i.e., *concatenate* and its inputs in Figure 3). Given a candidate term, query, and features, the neural network outputs the predicted term's predicted WRR.

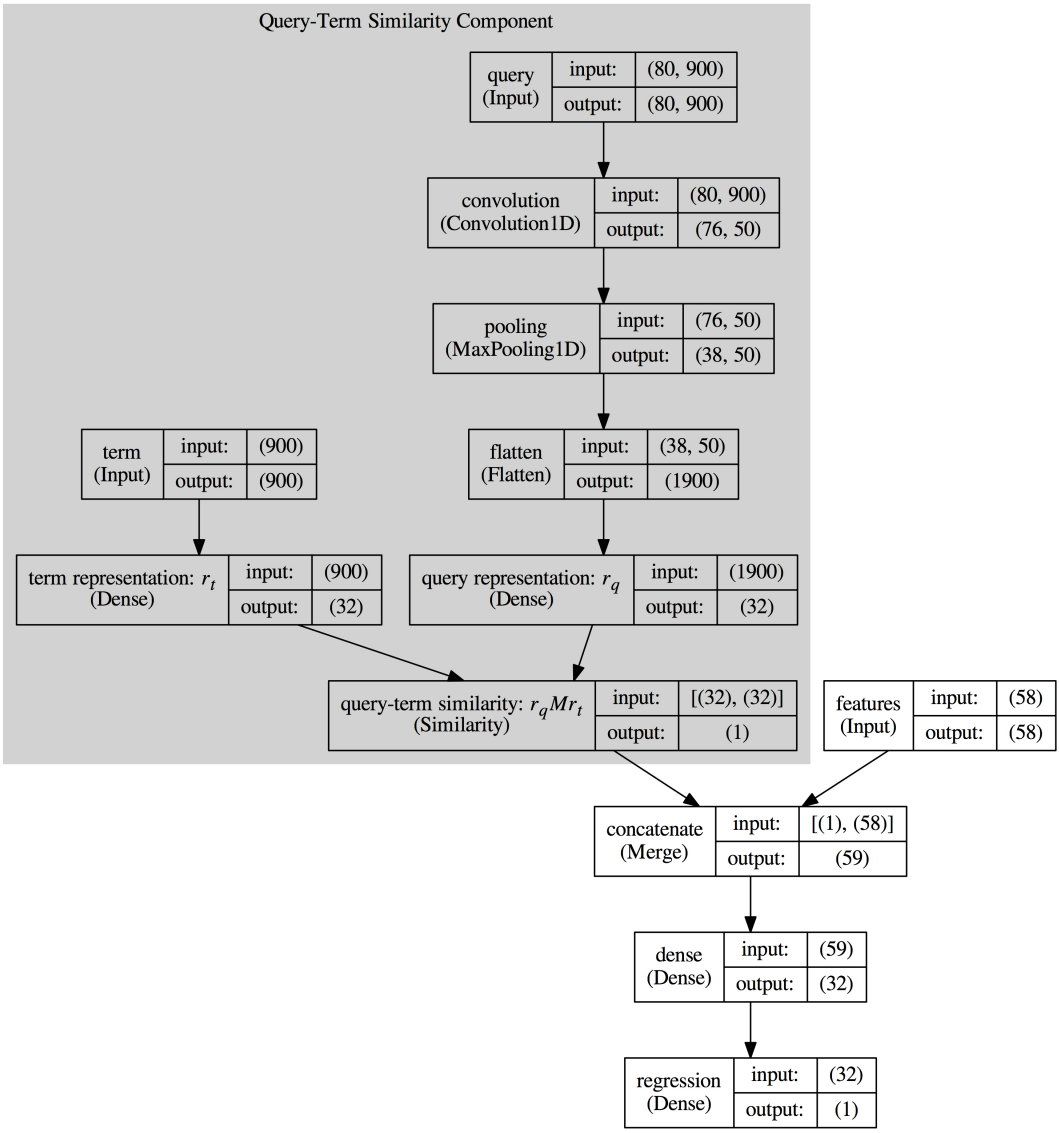

Figure 3: The Deep Neural Network (DNN) supervised candidate selection consists of a query-term similarity component and a feature component. Each square represents a layer. Arrows indicate each layer's input. Layer types are shown in parentheses; *Flatten* and *Merge* layers modify their inputs' shape without modifying the input itself. Query-term similarity is computed by the *query-term similarity* layer (shaded in gray in the figure) as described in the query-term similarity section, combined with other features in the *concatenate* layer, and input to two dense layers (i.e., *dense* and *regression*) to perform the regression based on the query-term similarity and the term's features.

Query-Term Similarity

Our query-term similarity component learns compact query and term representations and computes their similarity with the help of a learned similarity matrix M. That is,

$$sim(r_q, r_t) = r_q M r_t \tag{4}$$

where r_q and r_t are compact query and term representations, respectively.

We learn the query representation by using a 1-dimensional convolution over a `word2vec` representation of the query, and applying $n_{filters}$ filters to the convolution followed by max pooling and a dense layer with a ReLU activation function and $n_{representation}$ neurons (i.e., *query representation* in Figure 3). That is, the convolution layer combines each w-term sliding window with $n_{filters}$ filters to produce $n_{filters}$ features for each sliding window, before using max pooling to take the top 50% of query term sliding windows and creating a compact representation of the query.

The convolutional layer's purpose is to apply position-independent filters to w-term windows of query terms. Without the convolution, the query representation would be dependent on the exact position in the query each term appears in. The dense layer's purpose is to learn to reduce the dimensionality of the query representation; the representation vector must be small both to generalize from training to testing and to match the dimensionality of the term representation vector.

Similarly, we learn the term representation by feeding a `word2vec` representation of the term to a single dense layer with a ReLU activation function and $n_{representation}$ neurons (i.e., *term representation* in Figure 3). As with the query representation dense layer, the term representation dense layer's purpose is to learn to reduce the dimensionality of the term representation.

The output of these steps is a query representation vector r_q and a term representation vector r_t with $n_{representation}$ dimensions. Finally, we compute the similarity between r_q and r_t as described above (i.e., using *query-term similarity* in Figure 3) and pass $sim(r_q, r_t)$ to the neural networks second component (i.e., *concatenate* in Figure 3).

Features

Recently, Oh and Jung (2015) have shown that taking advantage of multiple document collections leads to significant improvements in medical literature retrieval. Similarly, we consider several collections of health documents to capture medical soundness of candidate terms, as well as relationships between expansion candidates and query terms. The following collections were used to obtain features for candidate terms:

- **Khreshmoi project**[8] (Hanbury et al., 2011): a collection of approximately 1.1 million web pages in the health domain. Pages in the collection were sampled from websites that have been certified by the HON foundation. Other known trustworthy websites were also included.
- **Health Wikipedia**: 22,943 Wikipedia pages from its Portal of Medicine[9]. This set of pages was extracted using the previously described information box heuristic.

[8] `http://www.khresmoi.eu/`

[9] `https://en.wikipedia.org/wiki/Portal:Medicine`

- **Wikipedia**: a set of 5.9 million English Wikipedia pages collected on May 5, 2016. While pages in this collection are not necessarily from the medical domain, it should help discerning medical terminology from general domain terms.
- **PubMed Central**: the open access subset of PubMed Central[10]. The snapshot we use — obtained on January 21, 2014 — is the same test collection used in the CDS track at TREC.
- **A.D.A.M. Medical Encyclopedia**: a consumer-oriented medical encyclopedia. We use the subset available through Medline Plus[11], which consists of 1,789 pages. This dataset was retrieved in May 2016.
- **MedScape**: a collection of 7,590 pages containing educational material (e.g., summaries of diseases, descriptions of symptoms, lists of drugs interactions, differential diagnosis sheets, etc.) for medical specialists, primary care physicians, and other health professionals. The collection was retrieved in June 2016.

For each collection $\mathcal{C}$ and each candidate term t, we consider the inverse document frequency (idf) of the term in the collection as a feature. Specifically, the following formulation of idf is used:

$$idf(t, \mathcal{C}) = \log_{10}\left(\frac{|\mathcal{C}| + 1}{df(t, \mathcal{C}) + 1}\right) \tag{5}$$

where $df(t, \mathcal{C})$ is the document frequency of term t in collection C, i.e., the number of documents in $\mathcal{C}$ that contain t.

To capture the semantic relationship between query terms and candidate terms, we extract, for each candidate term t, query term q, and collection $\mathcal{C}$, the number $N_{t,q,\mathcal{C}}$ of documents in which t and q co-occur; Then, for each t, we consider as feature the minimum, maximum, average, and standard deviation of $N_{t,q,\mathcal{C}}$ for all terms in the query.

Finally, similarly to (Soldaini and Goharian, 2017), we also consider the following features for each candidate term:

- The PRF score of the term, as defined in Equation 1.
- The odds ratio of the term, as defined in Equation 2.
- The number of concepts in the UMLS metathesaurus that can be matched to the candidate term; QuickUMLS (Soldaini and Goharian, 2016) was used to identify concepts.
- The number of concepts in UMLS that contain the candidate term; note that this differs from the previous features, as a term that is not a UMLS concept (e.g., "swine") can still appear as part of one ("african swine fever").
- The length in characters of the candidate terms.

[10] `https://www.ncbi.nlm.nih.gov/pmc/`
[11] `https://medlineplus.gov/encyclopedia.html`

Rank	Feature	ρ_s
1	HTPRF score	0.426
2	odds of being in health Wikipedia	0.134
3	term is a noun	0.095
4	term is a verb	-0.094
5	term is a UMLS concept	0.093
6	MedScape *co-occurrence st.dev.*	0.089
7	English Wikipedia *idf*	-0.083
8	MedScape *co-occurrence max.*	0.082
9	term is part of UMLS concept	-0.068
10	MedScape *co-occurrence avg.*	0.067
11	Khreshmoi *co-occurrence min.*	0.066
12	Khreshmoi *co-occurrence st.dev.*	0.064
13	Khreshmoi *co-occurrence max.*	0.061
14	Health Wikipedia *co-occurrence min.*	0.060
15	A.D.A.M. *co-occurrence st.dev.*	0.059
16	length of term	0.033

Table 1: Top 16 features ranked by the absolute value of their Spearman's rank correlation coefficient (ρ_s) with WRR. All correlations are statistically significant (Spearman's rank correlation coefficient, two-tailed, $p < 0.05$).

- The Part of Speech (PoS) of the candidate term (e.g., the candidate term is a noun, verb, adjective, etc.).

In Table 1, we report the top 16 features, as determined by the absolute value of their Spearman's rank correlation coefficient (ρ_s) with WRR. We choose Spearman's rank correlation because the target value WRR—as well as many of the features—is not normally distribuited (Shapiro-Wilk test, two-tailed, $p < 0.05$). All correlations reported in the table are statistically significant (two-tailed, $p < 0.05$).

We note the two top ranked features are the HTPRF score and the odds ratio of a term appearing in health Wikipedia, two features that are used by HTPRF to select terms for expansion. This implies that the improved HTPRF is a strong baseline for the supervised method. Interestingly, the rank correlation suggests that candidate terms that are nouns are more likely to appear in relevant search results ($\rho_s = 0.095$), while verbs are more likely to appear in non-relevant search results ($\rho_s = -0.094$). As expected, collections whose content is mainly health-related (MedScape, Khreshmoi, health Wikipedia, A.D.A.M.) all have positive correlation with WRR, while English Wikipedia — which includes pages over many domains — correlates negatively with WRR.

Query Reformulation

Both HTPRF and DNN can be used to expand the preprocessed query. For the former, terms are ranked by their score; then, the top m candidate terms are used for expansion. As expected, the value of m affects the performance of the algorithm, as we will later discuss. For DNN, the top 30 terms by predicted WRR are added to the query.

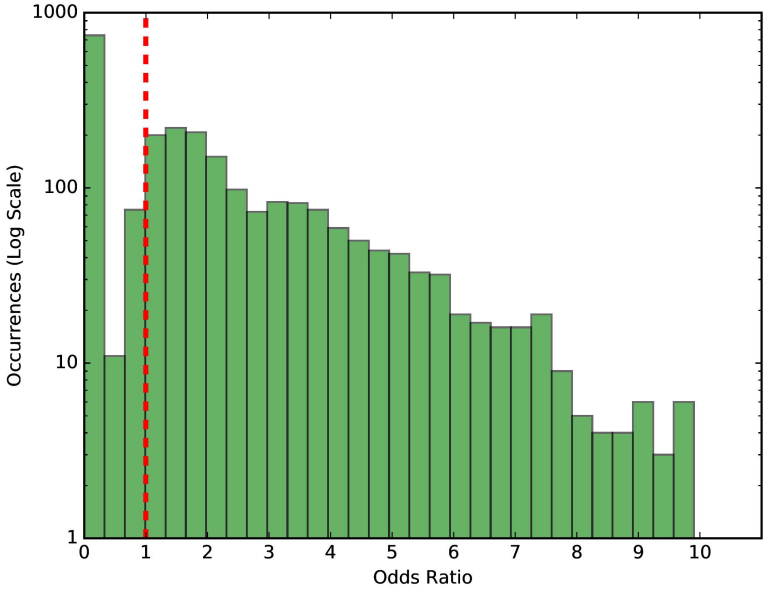

Figure 4: Distribution of the odds ratio of being relevant among terms in the query. Terms whose odds ratio is less than 1 (left of red dashed line) are more likely to appear in non-relevant documents than relevant documents. In our dataset, 832 query terms (34.6% of terms) have odds ratio less than 1.

As previously mentioned, queries for this task are long and discursive. Through statistical analysis, we determined that some terms in the queries are less likely to appear in relevant documents than others; thus, we experimented with query reduction algorithms to improve retrieval performance. In detail, the distribution of the odds ratio of terms is shown in Figure 4. Of 2,403 terms across 60 queries, 35% of them have an odds ratio less than 1, meaning that they are more likely to appear in non-relevant documents than in relevant documents. It follows that an effective query reformulation strategy that removes most of such terms would improve retrieval performances.

However, Soldaini et al. (2015) have shown that query reduction techniques that rely on extraction of UMLS concepts do not improve the performance of a CDS search system. Therefore, in this, work, we investigated whether part-of-speech (PoS) tags or syntactic dependencies could be used instead. We proceed as follows: first, we extract Part-of-Speech (PoS) tags and syntactic dependencies associated with the query. The two are coupled to identify all Noun Phrases (NP) in the query. The union of all noun phrases are considered as reformulated query. Furthermore, in (Soldaini et al., 2014), we suggested that Verb Phrases (VP) could have a significant impact in conveying the information need of each query. In this work, we set to study this by considering a query reduction algorithm that keeps both VPs and NPs.

To summarize, the following types of queries are expanded using the candidate terms as determined by the HTPRF and DNN:

Dataset year	Documents				Queries		Qrels	
	number	*has title*	*has abstract*	*has body*	*number*	*average length*	*relevant*	*non relevant*
2014	733,138	100%	86%	88%	30	78.6	3,356	34,594
2015						83.3	4,990	32,818

Table 2: Statistics of datasets used in the 2014 and 2015 CDS track at TREC. The same documents collection was used both years. "Qrels" is set of documents whose relevancy has been assessed by TREC organizers.

- Preprocessed query (stopwords, numbers, and units of measurement removed). We will refer to this method as "*stopwords removal*".
- Reduced preprocessed with terms t whose odds ratio of appearing in health Wikipedia is greater than or equal to δ (i.e., $\text{OR}(t) \geq \delta$). We will refer to this method as "*odds ratio reduction*".
- Reduced preprocessed query with only noun phrases. We will refer to this method as "*NP reduction*".
- Reduced preprocessed query with only noun phrases and verb phrases. We will refer to this method as "*NP+VP reduction*".

Dataset Description

The effectiveness of the proposed methods was studied on the datasets introduced in the CDS track at TREC 2014 (Roberts et al., 2016a) and TREC 2015 (Roberts et al., 2016b). The two dataset share the same documents collection, but have different sets of test queries. A summary of the two datasets is provided in Table 2.

Documents Collection

The collection of documents consists of a snapshot of the open access subset of PubMed Central (PMC). PMC is a database of biomedical literature; it is available online at no charge. The snapshot was defined by the organizers of the CDS track as the subset of all documents in PMC published before January 21, 2014. It contains 733,138 documents, totaling approximately 9.5 GB in size. Each article is in NXML format[12]. From each article, we extract the title, the abstract, and all sections in the body of the paper. Although all articles in PMC have a title, not all of them include a body or an abstract section. In the snapshot provided by the organizers of the CDS track, 14% of the articles have no abstract and 12% have no body. However, all articles have at least one of the two sections.

[12]NXML is a XML-compliant format whose tags are specified in the US National Library of Medicines Journal Archiving and Interchange Tag Library. A full specification is available at the following location: `http://jats.nlm.nih.gov/archiving/versions.html`.

Gobeill et al. (2014) computed the distribution of article types on the collection and on the set of relevant documents for the 2014 queries. Their work showed that 74.3% of articles in the collection are research articles (case reports: 4%; review articles: 6.9%; other: 15.8%). Similarly, 52.2% of relevant articles are research articles, 20.4% are case reports, and 17.9% are review articles. The remaining 9.5% belong to other categories.

Compared to our previous system (Soldaini et al., 2014, 2015), citation markers were removed using regular expressions; furthermore, we also removed table and figure captions. This more thorough preprocessing step is partially to credit for the improvements over our previous results.

Queries

The queries in both datasets were created by clinical informatics experts (all of whom were physicians) at the US National Library of Medicine. In the reminder of this section, we provide a brief description of them; we remand the reader to (Roberts et al., 2016a) for more details. Each query is comprised of three sections: a title, a summary, and a description. The description field was created to resemble a typical sign-out note (that is, a clinical note containing a brief history of a patient) in use at many hospitals when a patient is transfered across departments. In the words of the CDS track organizers, this process was done to "replicate the types of information contained in EHR notes, thus providing as near as possible a realistic evaluation of how such a retrieval system would perform in a clinical environment."

The information need of each query falls into one of these three categories: make a diagnosis, determine a test to confirm a diagnosis, establish the most appropriate treatment after diagnosis. We will refer to this categories as "diagnoses", "treatments", and "tests" throughout the rest of the manuscript. The three categories were chosen because previous research has shown that questions regarding diagnoses, treatments, and tests account for a 58% of the clinical questions posed by primary care physicians (Del Fiol et al., 2014). For each query, a summary and title were also provided. However, none of the proposed methods consider them for retrieval, as the description field is a more accurate representation of the search task studied in this work.

Experimental Setup

Our goals in designing an experimental plan was to quantify the difference between the improved HTPRF method, the DNN method described in this paper, and state-of-the-art proposed for the CDS search task. Furthermore, we were also interested in determining the effect of training parameters and features of the two methods proposed in this manuscript.

In order to carry out our experimental plan, we indexed the document collection using ElasticSearch[13]; Divergence from Randomness (Amati and Van Rijsbergen, 2002) was used as underlying similarity function.

For DNN, we train the neural network using the Adam algorithm (Kingma and Ba, 2014) for up to 30 epochs. Training is stopped early if loss fails to decrease on the validation set; in practice this happens after approximately 15 epochs. Term `word2vec` representations are obtained

[13] `https://www.elastic.co/products/elasticsearch`

HTPRF	DNN
anorexia **autonomic case diagnosis disorder distress episodes excessive** fatigue **gastrointestinal medication** nervosa **nocturnal onset patient psychiatric report restless severe signs sleep** symptoms syndrome tachycardia thyroid thyrotoxic thyrotoxicosis **treatment tremor**	anorexia **antithyroid emptying** fatigue **gastric graves hyperthyroidism hypoglycemia insomniacs meal methimazole milnacipran** nervosa **prandial propranolol remission remittent reuptake sertraline** symptoms syndrome tachycardia thyroid **thyroiditis** thyrotoxic thyrotoxicosis **triazolam**

Table 3: Example of terms added to the query shown in Figure 1 by the HTPRF (left) and DNN (right) methods. Terms in bold are exclusive to a method. For this query, HTPRF achieves higher P@10 (0.6 vs 0.3), but DNN achieves better infNDCG (0.419 vs 0.2506).

by concatenating 300-dimensional `word2vec` representations trained for 25 epochs on the PMC and Kreshmoi datasets described in the previous section; `word2vec` vectors are commonly 300 dimensions (Mikolov et al., 2013). In the neural network's second component, we use a dense layer with 32 neurons. Our implementation of DNN method leverages Gensim[14] and Theano[15]. Furthermore, spaCy[16] was used for PoS extraction.

In accordance with the CDS track at TREC, we consider inferred Normalized Discounted Cumulative Gain (infNDCG) as our primary metric. The choice was motivated by two reasons: first, it has been shown that inferred measures are capable of producing a more accurate estimate of the quality of a system when pooled judgments are used (Yilmaz et al., 2008); second, it allows for a direct comparison with systems that have participated to the track. Additionally, we also evaluate our system using precision at ten retrieved results (P@10). P@10 effectively estimates the ability of retrieving highly relevant, and thus actionable, medical literature in support of clinical practice, which is the goal of a CDS search system. We compare our system with the best teams at TREC 2014 (Mourao et al., 2013; Choi and Choi, 2014) and TREC 2015 (Balaneshin Kordan et al., 2015), as well as our previous system (Soldaini et al., 2014). A description of such systems was provided in the related works section.

Results and Analysis

In this section, we present an analysis of the performance of the two methods introduced in this manuscript. In detail, we first compare the proposed methods with the state of the art; then, we study the effect of query reduction techniques when combined with HTPRF and DNN query expansion methods; furthermore, we analyze the impact of individual features on the performance of the DNN method; finally, we detail the process of tuning parameters for HTPRF.

[14] `https://radimrehurek.com/gensim/`
[15] `http://deeplearning.net/software/theano/`
[16] `https://spacy.io/`

System	2014 dataset		2015 dataset	
	infNDCG	**P@10**	**infNDCG**	**P@10**
Baseline case report used as query	0.1546 -84.7%	0.2500 -56.0%	0.1729 -70.0%	0.3133 -54.2%
HTPRF (Soldaini et al., 2014)	0.2272 -24.7%	0.3200 -21.9%	0.2296 -28.0%	0.3367 -43.5%
SNUMedinfo (Choi and Choi, 2014)	0.2674 -5.9%	0.3633 -7.3%	n/a	n/a
NovaSearch (Mourao et al., 2014)	0.2631 -7.7%	**0.3900**	0.2242 -31.1%	0.3567 -35.5%
WSU-IR (Balaneshin Kordan et al., 2015)	n/a	n/a	**0.2939**	0.4667 -3.6%
improved HTPRF *(this work)*	*0.2567* *-10.3%*	*0.3733* *-4.5%*	*0.2653* *-10.8%*	***0.4833***
DNN expansion *(this work)*	***0.2833***	*0.3600* *-8.3%*	*0.2744* *-7.7%*	*0.4300* *-12.4%*

Table 4: Comparison of the proposed systems (last two rows) with a baseline method and the state of the art. For each column, the best result is in **bold**.

Comparison with State of the Art Systems

In Table 4, we report the performance of our methods on the 2014 and 2015 datasets. As previously mentioned, we compare the proposed approaches with the best approaches for the task, as well as with our previously proposed method. We also include a baseline system that uses the case report as query (no expansion; stopwords, numbers, and units of measurement removed). This baseline represents an important point of comparison with the two methods introduced in this work, since it is used as a first step to retrieve the top documents used to generate candidate terms for query expansion. We note that some results are missing due to the fact that some of the teams have not participated in both years.

Both systems proposed in this manuscript fare well against the state of the art. On the 2014 dataset, the DNN expansion approach outperforms any other method in terms of inferred NDCG, while NovaSearch achieves a better precision at 10. This behavior is expected, as NovaSearch uses a formulation of PRF in which expansion terms are chosen among high tf-idf terms in few top-ranked documents; this implicitly optimizes for precision at top ranked results. On the other hand, our DNN method is trained to choose terms based on WRR, which does not take into account their tf-idf score. On the 2015 dataset, the DNN method underperforms the state of the art, as well as the other method proposed in this manuscript, when measured by precision at 10.

The improved HTPRF method is also very competitive with respect to state of the art methods. The run reported in Table 4 uses odds ratio on Wikipedia to reduce the query before expanding it; a more detail analysis of query reduction is provided in a later section. Overall, we notice that, unlike the DNN expansion technique, HTPRF favors precision at 10 over inferred NDCG. This

could be a desirable characteristic of this method in those situations were obtaining a small set of highly relevant literature is preferred. On the 2014 dataset, HTPRF achieves a precision at 10 comparable to NovaSearch; on the 2015 dataset, it outperforms the state of the art, although WSU-IR achieves better infNDCG. We explain the substantial improvement in performance of HTPRF by observing that the baseline method — which is used to obtain the top k documents from which expansion terms are extracted — is also much more effective on the 2015 dataset, especially in terms of precision at top ranked results. This causes HTPRF to select more relevant terms from the top documents, which explains the increase in performance.

When comparing HTPRF with the DNN method, a few interesting observations can be made. First, we note that, for both methods, precision at 10 results and inferred NDCG strongly correlate (Pearson'r, $\rho = 0.7612$ for HTPRF, $\rho = 0.7885$ for supervised query expansion, $p < 0.05$ for both).

However, as shown in Figure 5, the relative performance of two methods vary depending on the query. In 25 out of 60 queries HTPRF outperforms the DNN method; the opposite occurs in the remaining 35 queries. On average, the DNN method outperforms HTPRF on diagnosis and tests, while the opposite happens for treatments. However, the difference is not statistically significant (Student t-test , two-tailed, $p = 0.83$, $p = 0.87$, and $p = 0.77$ respectively). Thus, we cannot conclude that the difference in infNDCG between the two methods is due to type of information need associated with the query.

Finally, we point out that the DNN method is more likely to choose UMLS concepts as expansion terms; on average, 82.3% of expansion terms selected by the DNN method are UMLS concepts, while only 72.5% of terms chosen by HTPRF are present in the metathesaurus (difference is statistically significant, Student t-test, two-tailed, $p < 0.05$). Using the semantic type associated with each concept and the taxonomy introduced in (Limsopatham et al., 2013), we were able to determine the aspects of the medical decision that the concepts chosen by the two methods belong to. For HTPRF, 18.5% of the terms are a diagnostic procedure or test (DNN: 19.3%), 17.1% are diseases (DNN: 19.7%), 32.5% are symptoms (DNN: 26.4%), and 20.3% are treatments (DNN: 23.4%). The remaining (11.6% and 11.2%) refer to other semantic types.

Impact of Query Reduction

Previous work (Soldaini et al., 2014, 2015) has suggested that query reduction could improve retrieval performance; therefore, we studied the impact of several query reduction techniques on the performance of both methods introduced in this manuscript. As previously mentioned, we set out to evaluate three query reduction techniques, and compared them with the original query (with stopwords, numbers, and units of measurement removed). Results for both methods are shown in Tables 5 and 6.

As shown in Table 5, removing terms that are less likely to appear in medical pages on Wikipedia is an effective strategy when combined with HTPRF, However, this technique is equally effective when combined with the DNN expansion method. We hypothesize that this is due to the fact that Equation 1 is likely to assign higher scores to terms that are semantically close to those in the query; thus, by removing less medically sound terms from the original query, we achieve an inprovement in inferred NDCG. Conversely, the DNN expansion method selects more diverse terms, thus increasing the need of keeping less medically sound terms in the query. This is evidenced by

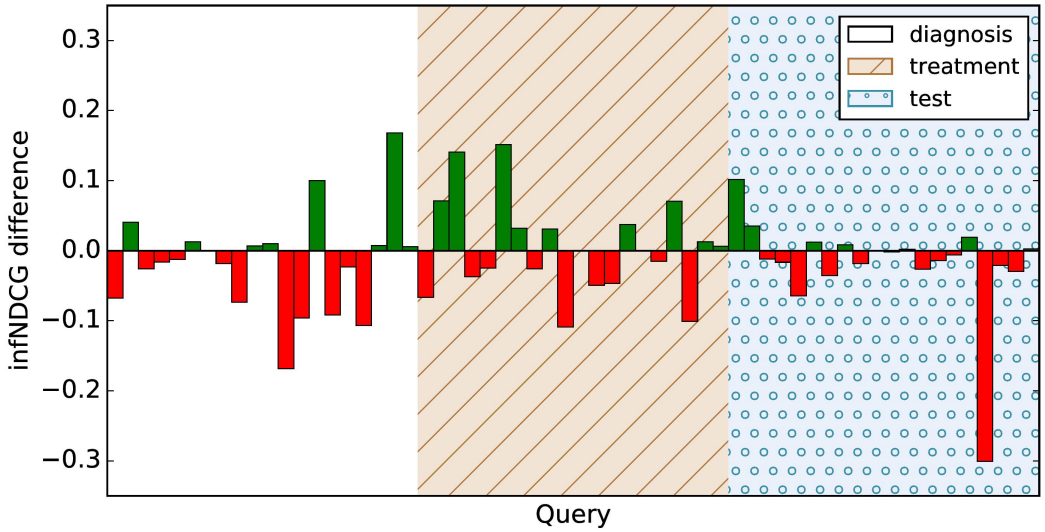

Figure 5: Difference in infNDCG between HTPRF and the DNN method for each query. Negative values are associated with queries in which the DNN method outperforms HTPRF.

the fact that average distance between UMLS concepts in the query and UMLS concepts in the candidate terms selected by HTPRF is 3.25 nodes in the UMLS graph, while the average distance for terms selected by the DNN method is 5.68 (difference is statistically significant, Student t-test, two-tailed, $p < 0.05$.)

Furthermore, we notice that, for the DNN expansion method, there exists a trade-off between infNDCG and P@10 when more aggressive query reduction algorithms are used (Table 5 and 6). "*NP reduction*" and "*NP+VP reduction*", which shorten the query substantially, cause an increase in inferred NDCG, but negatively affect precision at 10 retrieved results.

Overall, we note that query reduction techniques show limited improvement over the original method. As evidenced in Table 5 and 6, none of the reduction methods show statistically significant improvements in terms of infNDCG over simple stopwords removal (Student t-test, two-tailed, $p \geq 0.05$). The biggest improvements are with respect to P@10; when used with HTPRF, *odds ratio reduction*; however, it is less effective when paired with the supervised expansion method (DNN). When compared with the baseline, *odds ratio reduction* shows the best improvements on the 2014 dataset, while it performs similarly or worse on the 2015 dataset.

Impact of DNN Method Features

In this section we study the impact of different feature types on our DNN expansion method. To do so we individually evaluate (*i*) the query-term similarity component and feature component of our model, (*ii*) classes of features, and (*iii*) features derived from specific collections. The results are shown in Table 7. None of the changes in infNDCG and P@10 are statistically different from the model's performance with *DNN expansion: query-term similarity & features* (Student t-test,

System	2014 dataset		2015 dataset	
	infNDCG	**P@10**	**infNDCG**	**P@10**
improved HTPRF *stopword removal*	0.2541	0.3567	**0.2703**	0.4800
improved HTPRF *odds ratio reduction*	**0.2567**	**0.3733**	0.2653	**0.4833**
improved HTPRF *NP reduction*	0.2523	0.3633	0.2634	0.4367
improved HTPRF *NP+VP reduction*	0.2512	0.3533†	0.2621	0.4433*

Table 5: Comparison of several query reduction techniques on the improved HTPRF method. Query reduction using odds ratio achieves the best results except for a modest decrease in infNDCG on the 2015 dataset. However, the difference between runs is not statistically significant (Student t-test, two tailed, $p \geq 0.05$).

System	2014 dataset		2015 dataset	
	infNDCG	**P@10**	**infNDCG**	**P@10**
DNN expansion *stopwords removal*	0.2833	0.3600	0.2729	**0.4300**
DNN expansion *odds ratio reduction*	0.2842	**0.3700**	0.2698	0.4167
DNN expansion *NP reduction*	0.2865	0.3500	**0.2744**	0.4133
DNN expansion *NP+VP reduction*	**0.2919**	0.3400	0.2695	0.4267

Table 6: Comparison of several query reduction techniques on the DNN expansion method. *NP reduction* achieves the best infNDCG on the 2014 dataset, *NP+VP reduction* on the 2015 dataset, but both perform poorly in terms of P@10. Overall, the difference between runs is not statistically significant (Student t-test, two tailed, $p \geq 0.05$).

two-tailed, $p \geq 0.05$). This can be attributed to the low number of queries in our datasets. While excluding features can cause the average infNDCG and P@10 to change substantially, this change in the average metric is caused by substantial changes to a small number of queries. Over all the runs shown in Table 7, no more than 9 queries per run ever experience a change in infNDCG or P@10 greater than 0.1. The average number of queries experiencing such a change is much smaller; 3 queries for infNDCG and 7 queries for P@10 on the 2014 dataset, and 1 query for infNDCG and 4 queries for P@10 on the 2015 dataset. These values are much smaller than the number of queries for which P@10 changes in Tables 5 and 6: all runs that show a statistically significant difference experience a change in at least 13 out of 30 queries. We attribute this difference to the fact that query reduction methods potentially modify the entire expanded query, while the process of tuning the feature set for the supervised method only affects which new query expansion terms are added to the initial query.

System	2014 dataset		2015 dataset	
	infNDCG	**P@10**	**infNDCG**	**P@10**
DNN expansion *both (query-term sim. & features)*	0.2833	0.3600	0.2744	0.4300
DNN expansion *query-term similarity only*	0.2501	0.3100	0.2785	0.4200
DNN expansion *features only*	0.2726	0.3467	0.2714	0.4167
DNN expansion *both excluding IDF features*	0.2766	0.3033	0.2808	0.4400
DNN expansion *both excluding co-occurr. features*	0.2640	0.3600	0.2727	0.4233
DNN expansion *both excluding UMLS features*	0.2709	0.3633	0.2665	0.4167
DNN expansion *both excluding PRF features*	0.2545	0.3567	0.2785	0.4233
DNN expansion *both excluding odds ratio feature*	0.2762	0.3500	0.2761	0.4300
DNN expansion *both using only Wikipedia features*	0.2631	0.3567	0.2748	0.4100
DNN expansion *both using only A.D.A.M. features*	0.2606	0.3433	0.2767	0.4233
DNN expansion *both using only PubMed features*	0.2517	0.3567	0.2854	0.4233
DNN expansion *both using only MedScape features*	0.2627	0.3433	0.2691	0.4167

Table 7: Impact of model components, feature groups, and document collections on the DNN model's performance.

The model's performance using both components, only the query-term similarity component, and only the feature component are shown in the first three rows, respectively. While the 2015 infNDCGs are similar regardless of which components are used, using only the query-term similarity component substantially harms infNDCG and P@10 on the 2014 data set. The model performs better on the 2014 data when using only the feature component, but both components are necessary to achieve the best results.

The model's performance when different classes of features are excluded is shown in the next five rows of Table 7. The biggest change in performance as measured by infNDCG occurs when the UMLS features are excluded, causing the 2014 infNDCG to decrease from 0.2833 to 0.2709 and the 2015 infNDCG to decrease from 0.2744 to 0.2665. Excluding co-occurrence features and excluding PRF features both cause substantial decreases in performance on the 2014 data, but do not substantially affect the results on the 2015 dataset. Similarly, excluding the IDF features and excluding the odds ratio feature cause smaller decreases on the 2014 infNDCG, but slightly increase the 2015 infNDCG. We conclude that the UMLS features have the most impact on our model's

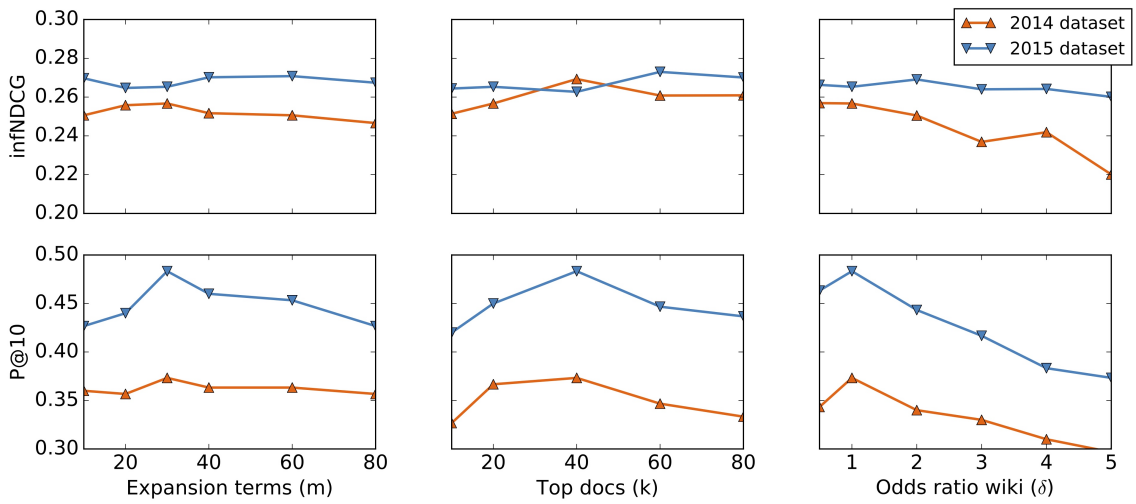

Figure 6: Effects of number of expansion terms (m, left), top documents (k, center), and minimum odds ratio (δ, right) on the performance of HTPRF, as measured by infNCCG (top) and P@10 (bottom). We chose $m = 30$, $k = 40$, $\delta = 1$.

performance, followed by the co-occurrence and PRF features.

Table 7's final four rows show the impact on performance when only features from specific collections are used (i.e., the co-occurrence and IDF features derived from a given collection). PubMed features perform the worst in terms of 2014 infNDCG, but perform the best in terms of 2015 infNDCG. The other three collections perform similarly, with MedScape performing slightly worse on 2015 infNDCG but not on 2014 infNDCG. This suggests that, when they are used independently, these three collections are somewhat interchangeable for the purpose of deriving co-occurrence and IDF features. The results improve when all the collections are used, however, suggesting that they are also complementary and it is beneficial to use multiple collections.

Parameter Tuning

Finally, in this section, we detail the tuning process we followed. In the case of HTPRF we chose the number of expansion terms m, the number of top documents k, and the minimum odds ratio δ for HTPRF. Our goal was to choose parameters that would maximize infNDCG across both datasets. The results of our optimization phase are shown in Figure 6. Alongside the effect of each parameter on infNDCG, we also present their effect on P@10.

We observed that HTPRF is moderately stable with respect to the choice of its parameters: even when varying m or k by two orders of magnitude, infNDCG was affected by at most 15%. However, HTPRF behaved differently between the two datasets. On the 2014 dataset, a smaller number of expansion terms and top documents achieves the best performances, while larger values of m and k were necessary to achieve better infNDCG on the 2015 dataset. Since queries in the

two datasets are of similar length and structure, we suspect that the differences in size of the pool of relevant documents (as shown in Table 2) might explain the different behavior: fewer relevant documents exists for the queries in the 2014 dataset. Thus, large values of k and m may cause query drift. Conversely, larger values of k and m are appropriate for the 2015 dataset, as more presumably relevant documents are considered to choose expansion candidates. Ultimately, because infNDCG is less sensitive to changes in k and m than P@10, we choose the values of k and m that maximize P@10; that is, we set $m = 30$ and $k = 40$. We stress that we did not intentionally choose the same parameters for the 2014 and 2015 datasets; rather, because of the heuristic described above, the two parameters set happen to be the same.

Contrary to k and m, the behavior of δ was consistent across the two dataset. Large values of δ caused too few terms to be selected for expansion, thus reducing the performance of HTPRF. Unlike infNDCG, P@10 behaves similarly across the two datasets when tuning parameters are varied.

To achieve a good balance between the two datasets, we chose the tuning parameters for our dataset by performing ten fold cross validation on the 2014 and 2015 datasets combined. In seven out of ten folds, parameters $m = 30$, $k = 40$, $\delta = 1$ maximized infNDCG; therefore, we chose such combination for all experiments reported in this section.

The DNN's parameters include the number of expansion terms m, the convolution size, the number of filters $n_{filters}$, and the term and query representation size $n_{representation}$. We found $m = 30$ terms to perform best on the 2014 dataset in terms of infNDCG. On the 2015 dataset varying the number of terms between 5, 10, 20, and 30 changed the average infNDCG by less than 1%. We thus used $m = 30$ terms in all experiments.

We empirically chose a convolution size of 5 (i.e., we consider 5 query terms at a time) with $n_{filters} = 50$ and $n_{representation} = 32$. Substantially increasing the number of filters (i.e., by more than 15%), the size of $n_{representation}$, or the dense layer harms performance by causing the neural network to overfit quickly, whereas substantially decreasing them reduces the network's ability to fit the training data and also harms performance. While there are many candidate terms to use as training data, the number of training queries is a limiting factor; additional training queries would likely allow these parameters to be increased.

Conclusions

In this work, we introduced two query reformulation techniques designed to address clinical decision support search, which is a search task intended to help medical professionals by retrieving medical literature that is pertinent to a given clinical note. Of the two systems, the first (HTPRF) is an improved version of unsupervised query expansion technique we introduced in a previous work. This approach combines pseudo relevance feedback with a health term filter designed to remove non-health related terms from the expansion candidates. The second method (DNN) is a supervised approach to query expansion based on a deep neural network for learning to rank short documents (Severyn and Moschitti, 2015); it leverages a deep neural network to predict each candidate terms weighted relevance ratio, a measure of importance of each term in relevant documents. To train the model, we use a combination of word embeddings, syntactical and semantic features over the candidate terms, and statistical features derived from the distribution of candidates and

query terms in several auxiliary collections.

The two approaches were evaluated on the CDS TREC 2014 and 2015 datasets. When compared to state of the art, the two systems fair well: on the 2014 dataset, DNN outperformed the state of the art by 7.7% in infNDCG; on the 2015, HTPRF outperformed existing systems by 3.6% in P@10; further analysis indicates that HTPRF generally exhibits better P@10, while DNN implicitly optimizes for infNDCG. Furthermore, we detail the tuning process for HTPRF and we study the impact of individual features for DNN.

Finally, we investigated query reduction approaches; first, we reasoned why query reduction might be an effective strategy for this task; then we presented three approaches to query reduction: terms removal based on their likelihood of being medical terms, noun phrase extraction, and noun and verb phrases extraction. We compared such methods to a simple stopwords removal baseline. Analysis of the performance of each method reveals that removing terms that are not frequently used in the medical domain improves the performance of HTPRF. Conversely, we saw less pronounced improvements for DNN.

Acknowledgments

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

References

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

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

Aronson, A. R. and Lang, F.-M. (2010). An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Balaneshin Kordan, S., Kotov, A., and Xisto, R. (2015). Wsu-ir at trec 2015 clinical decision support track: Joint weighting of explicit and latent medical query concepts from diverse sources. In *Proceedings of the 2015 Text Retrieval Conference*.

Bedrick, S. and Sheikshabbafghi, G. (2013). Lucene, metamap, and language modeling: Ohsu at clef ehealth 2013. In *CLEF (Working Notes)*.

Bendersky, M. and Croft, W. B. (2008). Discovering key concepts in verbose queries. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 491–498, New York, NY, USA. ACM.

Bendersky, M., Metzler, D., and Croft, W. B. (2010). Learning concept importance using a weighted dependence model. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 31–40, New York, NY, USA. ACM.

Burke, D. T., DeVito, M. C., Schneider, J. C., Julien, S., and Judelson, A. L. (2004). Reading habits of physical medicine and rehabilitation resident physicians. *American journal of physical medicine & rehabilitation*, 83(7):551–559.

Büttcher, S., Clarke, C. L., and Cormack, G. V. (2004). Domain-specific synonym expansion and validation for biomedical information retrieval (multitext experiments for trec 2004). In *TREC*.

Cao, Y., Xu, J., Liu, T.-Y., Li, H., Huang, Y., and Hon, H.-W. (2006). Adapting ranking svm to document retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 186–193. ACM.

Choi, S. and Choi, J. (2013). Snumedinfo at clefehealth2013 task 3. In *CLEF (Working Notes)*.

Choi, S. and Choi, J. (2014). Snumedinfo at trec cds track 2014: Medical case-based retrieval task. Technical report, DTIC Document.

Choi, S., Choi, J., Yoo, S., Kim, H., and Lee, Y. (2014). Semantic concept-enriched dependence model for medical information retrieval. *Journal of biomedical informatics*, 47:18–27.

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

Cormack, G. V., Clarke, C. L., and Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759. ACM.

de Herrera, A. G. S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., and Müller, H. (2013). Overview of the imageclef 2013 medical tasks. In *CLEF (Working Notes)*.

Del Fiol, G., Workman, T. E., and Gorman, P. N. (2014). Clinical questions raised by clinicians at the point of care: a systematic review. *JAMA internal medicine*, 174(5):710–718.

Diaz, F. and Metzler, D. (2006). Improving the estimation of relevance models using large external corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161. ACM.

Dong, L., Srimani, P. K., and Wang, J. Z. (2011). Ontology graph based query expansion for biomedical information retrieval. In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, pages 488–493. IEEE.

Duh, K. and Kirchhoff, K. (2008). Learning to rank with partially-labeled data. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 251–258. ACM.

Genkin, A., Lewis, D. D., and Madigan, D. (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304.

Gobeill, J., Gaudinat, A., Pasche, E., and Ruch, P. (2014). Full-texts representations with medical subject headings, and co-citations network reranking strategies for trec 2014 clinical decision support track. Technical report, DTIC Document.

Goeuriot, L., Jones, G. J., Kelly, L., Leveling, J., Hanbury, A., Müller, H., Salanterä, S., Suominen, H., and Zuccon, G. (2013). ShArE/CLEF eHealth evaluation lab 2013, task 3: Information retrieval to address patients' questions when reading clinical reports.

Goeuriot, L., Kelly, L., Li, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Jones, G., and Mueller, H. (2014). ShArE/CLEF eHealth evaluation lab 2014, task 3: User-centred health information retrieval. In *Proceedings of CLEF*, volume 2014.

Grossman, D. A. and Frieder, O. (2012). *Information Retrieval: Algorithms and Heuristics*. Springer.

Han, E.-H. S. and Karypis, G. (2000). Centroid-based document classification: Analysis and experimental results. In *European conference on principles of data mining and knowledge discovery*, pages 424–431. Springer.

Hanbury, A., Boyer, C., Gschwandtner, M., and Müller, H. (2011). KHRESMOI: towards a multilingual search and access system for biomedical information. *Med-e-Tel, Luxembourg*, 2011:412–416.

Haveliwala, T. H. (2002). Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM.

Hersh, W., Buckley, C., Leone, T., and Hickam, D. (1994). OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *SIGIR94*, pages 192–201. Springer.

Hersh, W., Price, S., and Donohoe, L. (2000). Assessing thesaurus-based query expansion using the UMLS metathesaurus. In *Proceedings of the AMIA Symposium*, page 344. American Medical Informatics Association.

Hersh, W. and Voorhees, E. (2009). Trec genomics special issue overview. *Information Retrieval*, 12(1):1–15.

Jalali, V. and Borujerdi, M. R. M. (2011). Information retrieval with concept-based pseudo-relevance feedback in MEDLINE. *Knowledge and information systems*, 29(1):237–248.

Jiang, J., Zheng, J., Zhao, C., Su, J., Guan, Y., and Yu, Q. (2016). Clinical-decision support based on medical literature: A complex network approach. *Physica A: Statistical Mechanics and its Applications*, 459:42–54.

Kalpathy-Cramer, J., de Herrera, A. G. S., Demner-Fushman, D., Antani, S., Bedrick, S., and Müller, H. (2015). Evaluating performance of biomedical image retrieval systemsAn overview of the medical image retrieval task at ImageCLEF 2004–2013. *Computerized Medical Imaging and Graphics*, 39:55–61.

Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., García Seco de Herrera, A., and Tsikrika, T. (2011). The CLEF 2011 medical image retrieval and classification tasks. In *Working Notes of CLEF 2011 (Cross Language Evaluation Forum)*.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kitanovski, I., Dimitrovski, I., and Loskovska, S. (2013). Fcse at medical tasks of imageclef 2013. In *CLEF (Working Notes)*.

Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., and Lawley, M. (2016). Information retrieval as semantic inference: a graph inference model applied to medical search. *Information Retrieval Journal*, 19(1):6–37.

Kumaran, G. and Carvalho, V. R. (2009). Reducing long queries using query quality predictors. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 564–571, New York, NY, USA. ACM.

Limsopatham, N., Macdonald, C., and Ounis, I. (2013). Inferring conceptual relationships to improve medical records search. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 1–8, Paris, France, France. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

Liu, T.-Y., Xu, J., Qin, T., Xiong, W., and Li, H. (2007). Letor: Benchmark dataset for research on learning to rank for information retrieval. In *Proceedings of SIGIR 2007 workshop on learning to rank for information retrieval*, pages 3–10.

Liu, Z. and Chu, W. W. (2007). Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Information Retrieval*, 10(2):173–202.

Lu, Z., Kim, W., and Wilbur, W. J. (2009). Evaluation of query expansion using mesh in pubmed. *Information retrieval*, 12(1):69–80.

Lv, Y. and Zhai, C. (2011). When documents are very long, BM25 fails! In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1103–1104. ACM.

Matos, S., Arrais, J. P., Maia-Rodrigues, J., and Oliveira, J. L. (2010). Concept-based query expansion for retrieving gene related publications from medline. *BMC bioinformatics*, 11(1):1.

McNamee, P. (2015). A domain independent approach to clinical decision support.

McNamee, P. and Mayfield, J. (2004). Character n-gram tokenization for european language text retrieval. *Information retrieval*, 7(1-2):73–97.

Mengle, S. S. and Goharian, N. (2009). Ambiguity measure feature-selection algorithm. *Journal of the American Society for Information Science and Technology*, 60(5):1037–1050.

Metzler, D. and Croft, W. B. (2005). A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479. ACM.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

Mourao, A., Martins, F., and Magalhaes, J. (2013). Novasearch on medical imageclef 2013. In *CLEF (Working Notes)*.

Mourao, A., Martins, F., and Magalhaes, J. (2014). Novasearch at trec 2014 clinical decision support track. Technical report, DTIC Document.

Müller, H., de Herrera, A. G. S., Kalpathy-Cramer, J., Demner-Fushman, D., Antani, S., and Eggel, I. (2012). Overview of the imageclef 2012 medical image retrieval and classification tasks. In *CLEF (online working notes/labs/workshop)*, pages 1–16.

Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Bakke, B., Kahn Jr, C. E., and Hersh, W. (2009). Overview of the clef 2009 medical image retrieval track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 72–84. Springer.

Müller, H., Kalpathy-Cramer, J., Eggel, I., Bedrick, S., Radhouani, S., Reisetter, J., Kahn Jr, C. E., and Hersh, W. (2010). Overview of the clef 2010 medical image retrieval track. In *Working Notes of CLEF 2010 (Cross Language Evaluation Forum)*. Springer.

Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Frnkranz, J. and Joachims, T., editors, *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814. Omnipress.

Naveh, E., Katz-Navon, T., and Stern, Z. (2015). Resident physicians clinical training and error rate: the roles of autonomy, consultation, and familiarity with the literature. *Advances in Health Sciences Education*, 20(1):59–71.

Oh, H.-S. and Jung, Y. (2014). A multiple-stage approach to re-ranking clinical documents. In *CLEF (Working Notes)*, pages 210–219.

Oh, H.-S. and Jung, Y. (2015). Cluster-based query expansion using external collections in medical information retrieval. *Journal of biomedical informatics*, 58:70–79.

Roberts, K., Simpson, M., Demner-Fushman, D., Voorhees, E., and Hersh, W. (2016a). State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the trec 2014 cds track. *Information Retrieval Journal*, 19(1-2):113–148.

Roberts, K., Simpson, M. S., Voorhees, E., and Hersh, W. R. (2016b). Overview of the trec 2015 clinical decision support track.

Rocchio, J. J. (1971). Relevance feedback in information retrieval.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Sankhavara, J., Thakrar, F., Sarkar, S., and Majumder, P. (2014). Fusing manual and machine feedback in biomedical domain. Technical report, DTIC Document.

Severyn, A. and Moschitti, A. (2015). Learning to rank short text pairs with convolutional deep neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 373–382, New York, NY, USA. ACM.

Shen, W., Nie, J.-Y., Liu, X., and Liui, X. (2014). An investigation of the effectiveness of concept-based approach in medical information retrieval grium at clef2014ehealthtask 3. *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*.

Sierek, T. and Hanbury, A. (2015). Using health statistics to improve medical and health search. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 287–292. Springer.

Simpson, M. S., You, D., Rahman, M. M., Demner-Fushman, D., Antani, S., and Thoma, G. R. (2013). Iti's participation in the 2013 medical track of imageclef. In *CLEF (Working Notes)*. Citeseer.

Soldaini, L., Cohan, A., Yates, A., Goharian, N., and Frieder, O. (2014). Query reformulation for clinical decision support search. In *23rd Text REtrieval Conference (TREC)*.

Soldaini, L., Cohan, A., Yates, A., Goharian, N., and Frieder, O. (2015). Retrieving medical literature for clinical decision support. In *European Conference on Information Retrieval (ECIR)*, pages 538–549. Springer.

Soldaini, L. and Goharian, N. (2016). QuickUMLS: a fast, unsupervised approach for medical concept extraction. In *Proceedings of the 2nd Medical Information Workshop (MedIR) at the 34th international ACM SIGIR conference on Research and development in Information Retrieval*.

Soldaini, L. and Goharian, N. (2017). Learning to rank for consumer health search: a semantic approach. In *European Conference on Information Retrieval (ECIR)*. Springer.

Srinivasan, P. (1996). Query expansion and MEDLINE. *Information Processing & Management*, 32(4):431–443.

Stokes, N., Li, Y., Cavedon, L., and Zobel, J. (2009). Exploring criteria for successful query expansion in the genomic domain. *Information retrieval*, 12(1):17–50.

Tenopir, C., King, D. W., Clarke, M. T., Na, K., and Zhou, X. (2007). Journal reading patterns and preferences of pediatricians. *Journal of the Medical Library Association*, 95(1):56.

Thesprasith, O. and Jaruskulchai, C. (2014). Query expansion using medical subject headings terms in the biomedical documents. In *Asian Conference on Intelligent Information and Database Systems*, pages 93–102. Springer.

Urbain, J., Frieder, O., and Goharian, N. (2009). Passage relevance models for genomics search. *BMC bioinformatics*, 10(3):S3.

Voorhees, E. M. (2013). The trec medical records track. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, BCB'13, pages 239:239–239:246, New York, NY, USA. ACM.

Voorhees, E. M. and Hersh, W. R. (2012). Overview of the trec 2012 medical records track. In *TREC*.

Voorhees, E. M. and Tong, R. M. (2011). Overview of the trec 2011 medical records track. In *TREC*.

Xu, J. and Li, H. (2007). Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 391–398. ACM.

Xu, T., McNamee, P., and Oard, D. W. (2014). Hltcoe at trec 2014: Microblog and clinical decision support.

Yilmaz, E., Kanoulas, E., and Aslam, J. A. (2008). A simple and efficient sampling method for estimating ap and ndcg. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 603–610. ACM.

Zhai, C. and Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM.

Zheng, Z., Brady, S., Garg, A., and Shatkay, H. (2005). Applying probabilistic thematic clustering for classification in the trec 2005 genomics track. In *TREC*.

Zhu, D., Wu, S., Carterette, B., and Liu, H. (2014). Using large clinical corpora for query expansion in text-based cohort identification. *Journal of biomedical informatics*, 49:275–281.

Zhu, D., Wu, S. T.-I., Masanz, J. J., Carterette, B., and Liu, H. (2013). Using discharge summaries to improve information retrieval in clinical domain. In *CLEF (Working Notes)*.