Optimization Method for Weighting Explicit and Latent Concepts in Clinical Decision Support Queries

Saeid Balaneshin-kordan
Department of Computer Science
Wayne State University
Detroit, Michigan 48202
saeid.balaneshinkordan@wayne.edu

Alexander Kotov
Department of Computer Science
Wayne State University
Detroit, Michigan 48202
kotov@wayne.edu

ABSTRACT

Accurately answering verbose queries that describe a clinical case and aim at finding articles in a collection of medical literature requires capturing many explicit and latent aspects of complex information needs underlying such queries. Proper representation of these aspects often requires query analysis to identify the most important query concepts as well as query transformation by adding new concepts to a query, which can be extracted from the top retrieved documents or medical knowledge bases. Traditionally, query analysis and expansion have been done separately. In this paper, we propose a method for representing verbose domain-specific queries based on weighted unigram, bigram, and multi-term concepts in the query itself, as well as extracted from the top retrieved documents and external knowledge bases. We also propose a graduated non-convexity optimization framework, which allows to unify query analysis and expansion by jointly determining the importance weights for the query and expansion concepts depending on their type and source. Experiments using a collection of PubMed articles and TREC Clinical Decision Support (CDS) track queries indicate that applying our proposed method results in significant improvement of retrieval accuracy over state-of-the-art methods for ad hoc and medical IR.

CCS Concepts

•**Information systems** → **Query reformulation;**

Keywords

Medical Literature Retrieval; Clinical Decision Support; Feature-based Retrieval Models; Optimization Methods

1. INTRODUCTION

Given descriptive summary of a medical case as a query, the goal of information retrieval systems for clinical decision support (CDS) is to return articles from a collection of medical literature that are relevant to the query and can assist a clinician in making decisions regarding the case, such as prescribing a medication, procedure or treatment. A fundamental challenge faced by those systems is that although CDS queries are typically verbose and may consist of several sentences (e.g. *"33-year-old male presents with severe abdominal pain one week after a bike accident, in which he sustained abdominal trauma. He is hypotensive and tachycardic, and imaging reveals a ruptured spleen and intraperitoneal hemorrhage"*), only a small subset of query terms (henceforth referred to as explicit concepts) correspond to the key query concepts, such as *"bike accident"*, *"abdominal trauma"*, *"tachycardia"*, *"splenic rupture"*, *"intraperitoneal hemorrhage"*, which represent the information need behind this query, while many other important concepts that are relevant to this information need (e.g. *"spontaneous spleen rupture"*, *"splenic trauma"*, etc.) are not directly mentioned in the query (henceforth referred to as latent concepts). Providing complete and accurate retrieval results for CDS queries requires both correct identification of the key explicit concepts and addition of important latent concepts to the query, as well as precise weighting of explicit and latent concepts in the modified query.

While previous work on general and domain-specific IR has focused on identification of the key statistical concepts in verbose queries [3, 4, 5], latent query concepts in external resources ([14, 30, 37, 38]) and the top-retrieved (PRF) documents [5, 17] individually, to the best of our knowledge, no query transformation method that uses both explicit concepts from the query and latent concepts from diverse sources, such as external resources and PRF documents, has been previously proposed. For example, Latent Concept Expansion (LCE) [17] and Parameterized Query Expansion (PQE) [5] methods use only unigrams from the top-retrieved documents as latent concepts, while [11] uses only unigrams from structured knowledge bases as latent concepts for query expansion.

In this work, we propose a novel method to represent verbose clinical decision support queries using unigram, bigram and multi-term concepts from the query itself, as well as from the PRF documents and external knowledge bases (such as the Unified Medical Language System). Our method is based on linear feature-based learning-to-rank retrieval framework [18], in which the relative importance weight is determined for each matching query concept individually as a linear combination of features. We also propose a set of features for each concept type, which is determined based on whether a concept is a unigram, bigram or multi-term phrase

ICTIR '16, September 12-16, 2016, Newark, DE, USA

DOI: http://dx.doi.org/10.1145/2970398.2970418

and whether it occurs in the query itself or is extracted from a top retrieved document or a knowledge base.

Since the parameter spaces of linear feature-based retrieval models can be reduced to a multinomial manifold, their parameters can be estimated by direct maximization of the target rank-based retrieval metric (e.g. NDCG) over this manifold using derivative-free unconstrained multi-dimensional optimization methods, such as coordinate ascent [19] or hill-climbing [21]. These methods are based on the Powell's method, which divides a complex multi-dimensional optimization problem into several simple one-dimensional ones. After that, it iteratively optimizes a multivariate objective function by optimizing each parameter individually, while holding all other parameters fixed. Since line search is a local optimization method, the efficiency and accuracy of both the coordinate ascent and hill-climbing rely on the assumption of smoothness and convexity of objective function when a free parameter is optimized, which is often violated in practice. Figure 1, which shows the behavior of the target retrieval metric by varying the value of a parameter that corresponds to the weight of a feature, illustrates this case. It can be seen that the objective function shown in this figure has several local maxima.

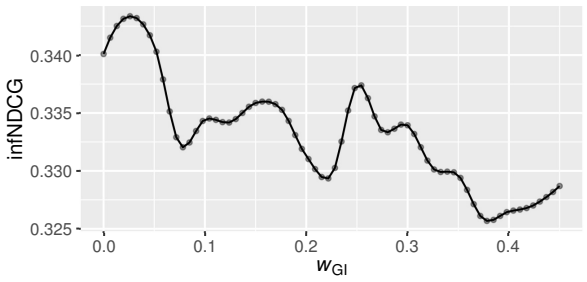


Figure 1: The values of objective function corresponding to infNDCG retrieval metric by varying the weight of one of the features (GI presented in Section 3.3), which determines the importance of concept matches of certain type.

The optimization method for learning the weights of concept importance features in feature-based retrieval models proposed in this paper leverages the Graduated Non-Convexity (GNC) (or continuation) optimization method [6] to address the issue of non-smooth and non-convex objective functions, when individual parameters are optimized using the Powell's method. GNC is a derivative-free method specifically designed for global optimization of non-smooth and non-convex objective functions. Graduated Non-Convexity (GNC) is an iterative method, which applies different degrees of smoothing to the original objective function to generate smoother and more convex objective functions, which have their global maximum close to the one of the original objective function. The method starts by applying the highest degree of smoothing and then gradually decreases the rate of smoothing at each subsequent iteration using the result obtained at the previous iteration as the starting point for the next iteration until the global maximum for the original non-smoothed objective function is found. Although the quality of the solution attained by this approach heavily depends on the choice of the smoothing method, it was recently shown that Gaussian smoothing of a non-convex function is optimal in a sense that it evolves any function into its convex envelope [20].

The remainder of this paper is organized as follows. After a brief summary of related work in Section 2, we discuss the details of the ranking function, the features and the optimization method to estimate the concept importance weights in Section 3. Section 4 provides the results of an experimental evaluation of retrieval accuracy of the proposed method with respect to the state-of-the-art baselines, while Section 5 concludes the paper.

2. RELATED-WORK

Depending on the type of concepts used for query expansion, general-purpose and domain-specific retrieval methods can be categorized into the ones that are based on statistical concepts (i.e. determined based on term popularity and co-occurrence in a given collection) [3, 5, 16, 17, 31], the ones that are based on semantic concepts (i.e. that are extracted from a knowledge repository) [13, 29, 30, 37], and those that combine semantic and statistical concepts [7, 25, 34, 9]. Below we provide an overview of the previously proposed methods in each of these 3 categories.

Retrieval methods using statistical concepts. In the simplest case, these retrieval models utilize only unigrams from the top retrieved documents for query expansion [31]. More recent retrieval methods utilizing statistical concepts are based on the Markov Random Field (MRF) framework introduced by Metzler and Croft [16]. It assigns the same importance weight to all matching statistical query concepts of the same type (unigrams and sequential bigrams), when the retrieval score of a document is calculated. Latent Concept Expansion (LCE) extends MRF by also using unigrams from the PRF documents as latent concepts for query expansion. The requirement of having fixed weights for unigrams and bigram concepts in the MRF-based retrieval model was relaxed by the Weighted Sequential Dependence (WSD) model [4], which estimates the importance of each concept individually. A similar relaxation of LCE weights was implemented in the Parameterized Query Expansion (PQE) [5] model. Overall, query representation methods based on statistical concepts typically consider unigrams and bigrams in the query and/or unigrams in PRF documents.

Retrieval methods using semantic concepts. Semantic concepts for query expansion are typically extracted from domain-specific, such as the Unified Medical Language System (UMLS) [13], Medical Subject Headings (MeSH) [15] and Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) [10], or general-purpose knowledge repositories, such as Wikipedia [29, 35]. The utility of this type of concepts has been studied for a variety of medical IR tasks including medical literature retrieval [27, 38]. UMLS concepts are typically extracted from queries and top-retrieved documents using MetaMap [1, 8, 13, 28, 29, 32, 33].

Soldaini et al. [29] proposed two methods for medical literature retrieval that use Wikipedia-based heuristics to filter out non-medical concepts from the original query and top retrieved documents. The first method (referred to as `HT` in [29] and `Wiki-Orig` in this work) is a query reduction method, which retains only those bigram concepts in the original query that are determined to be health-related according to a heuristic. On the other hand, the second method (referred to as `HT-PRF` in [29] and `Wiki-TD` in this

work) expands the original query with a number of health-related concepts that are extracted from the top-retrieved documents and filtered out using the same heuristic.

Accounting for semantic types of concepts[1] can also significantly improve the accuracy of query expansion, as they can be used to filter out the candidate expansion concepts. The method proposed in [13] (referred to as `UMLS-TD` in this work), expands medical queries only with the UMLS concepts extracted from the top retrieved documents that have pre-selected semantic types. A semantic type is pre-selected if the concepts of this type improve the accuracy of retrieval results when added to the queries in the training set. For example, the semantic type "signs and symptoms" is pre-selected for a query about the diagnosis of a disease. [32] proposed another approach to using semantic types in the query, in which the semantic types of concepts are used to weight the concepts (concepts that are more likely to be effective, get higher weight).

Retrieval models using both semantic and statistical concepts. The benefit of integrating semantic and statistical concepts was shown in [7, 9, 25, 34]. The methods in [7, 25, 34] focused only on explicit concepts (query unigrams and bigrams along with UMLS concepts extracted from the query using MetaMap). A medical IR system that integrates a graph-based representation of the corpus, structured knowledge sources and a retrieval model combining statistical IR methods with an inference mechanism implemented as graph traversal has been proposed in [9].

The key difference of the proposed method from existing methods for medical literature and ad hoc document retrieval is that it uses both statistical and semantic concepts extracted from diverse sources (query itself, knowledge bases and top retrieved documents) for query representation. The proposed method also leverages an efficient optimization technique to learn the relative importance weight of different types of query concepts on the same scale.

3. METHOD

In this section, we present the details of the proposed query reformulation method, a set of features used with it and a method to optimize the weights of those features with respect to the target retrieval metric. The proposed query reformulation method combines explicit and latent query concepts from diverse sources and determines the weight of each individual concept as a linear combination of features, which depend on a concept type. The type of a query concept is determined by its source and whether the concept is represented by a unigram, bigram or multi-word phase. The set of concept sources considered in our method includes the query itself, top retrieved documents for the original query, and external knowledge repositories.

3.1 Retrieval model

To account for term dependencies, the proposed method adopts a Markov Random Field (MRF) retrieval framework [16], in which the retrieval score of a document is determined as a weighted linear combination of the matching scores of different concept types in a given query. In particular, our method extends the parametrized concept retrieval model in [5], according to which the retrieval score of document D with respect to query Q is calculated as:

$$sc(Q,D) = \sum_{T \in \mathcal{T}_Q} \sum_{c \in \mathcal{C}_T} \lambda_T(c) f_T(c,D) \quad (1)$$

where $\mathcal{C}_T$ is a set of concepts belonging to concept type T, and $\lambda_T(c)$ is defined as the importance weight of concept c, which depends on its type. In the above equation, $f_T(c,D)$ is the matching score function of concept c in document D, which is defined as:

$$f_T(c,D) = \log((1-\lambda)\frac{n(c,D) + \mu\frac{n(c,Col)}{|Col|}}{|D|+\mu} + \lambda\frac{n(c,Col)}{|Col|}) \quad (2)$$

where $n(c,D)$ ($n(c,Col)$) and $|D|$ ($|Col|$) are the counts of concept c in document D (entire collection) and the size of document D (entire collection), respectively. The above matching function utilizes a two-stage smoothing method from [36], where λ and μ are Jelinek-Mercer and Dirichlet smoothing coefficients, respectively. Since only unigrams as well as ordered and unordered bigrams are considered in the MRF retrieval framework, concepts that are represented by multi-word phrases are broken down into unigrams and sequential bigrams. The set of concept types considered for a query Q is designated by $\mathcal{T}_Q$ and is shown in Table 2. This table also provides information about the concept extraction methods and a set of features corresponding to each concept type, which will be explained in detail below.

The importance weight of concept c is parameterized using a set of importance features $\Phi_T(c)$. Each concept type T is associated with its own set of importance features, summarized in Table 1. Thus, the weight of concept c with type T is determined as a weighted linear combination of importance features:

$$\lambda_T(c) = \sum_{n=1}^{N} w_\phi^n \phi_n \ , \quad (3)$$

where $\{\phi_1, \ldots, \phi_N\}$ is a set of features for concepts with type T (i.e., $\Phi_T(c) = \{\phi_1, \ldots, \phi_N\}$), and w_ϕ^n is the importance weight of the n-th feature (i.e., ϕ_n). The intuition behind this concept weighting scheme is that different concept types have different importance and should be weighted accordingly. Intuitively, knowledge-based concepts (such as the UMLS concepts) that are linked from the concepts in the original query should have a different importance weight than the concepts that are extracted from the top retrieved documents. Similarly, bigrams corresponding to UMLS concepts identified in the original query should be weighted differently than other bigrams in the original query. On the other hand, features determining the importance of a concept from a graph structured knowledge repository (e.g. UMLS), like the degree of the node corresponding to this concept, are different from the features that determine the importance of a unigram concept in top retrieved documents.

3.2 Optimization Method

Learning the feature weights that maximize the target retrieval metric on a training data can be considered as a multivariate optimization problem and is typically addressed by decomposing it into a set of one-dimensional optimization problems. Instead of performing a line search along every single dimension in optimizing a set of feature weights with respect to the target retrieval metric, we propose to use grad-

[1]http://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml

uated optimization [6], an efficient global optimization technique.

3.2.1 Graduated optimization

Graduated optimization is an iterative optimization method that gradually finds the global optimum of a given objective function by finding the optima for a series of simplified objective functions. Each of these simplified objective functions is obtained from the original objective function by applying different degree of smoothing to make the original function more convex. It starts from the solution to the most simplified optimization problem (i.e., when the maximum degree of smoothing is applied to the original objective function) and considers this solution as the starting point for the second less simplified problem (i.e. less smoothed original objective function). This process continues until the global optimum for the original objective function is found. This procedure is based on the assumption that the global optimum of a given objective function at the current iteration is close enough to its global optimum at the next iteration. Therefore, at the next iteration, the region of the parameter space that is far enough from the optimum point at the current iteration is ignored. As a result, a smaller region that is close to the optimum point at the current iteration is searched for the optimal parameter setting at the next iteration.

3.2.2 Smoothing method

In case of a univariate optimization problem with a single parameter w_ϕ, the smoothed objective function, $\tilde{E}(w_\phi)$, can be obtained by taking sample values from $E(w_\phi)$, the original objective function. To compute $\tilde{E}(w_\phi)$ at a specific region around the starting point $w_{\phi,0}$, samples are taken from $\tilde{E}(w_\phi)$ for the following values of w_ϕ:

$$\mathbf{w}_{s,\phi} = [w_{\phi,-M}, \ldots, w_{\phi,0}, \ldots, w_{\phi,M}] \tag{4}$$

where

$$w_{\phi,m} = w_{\phi,0} + m\Delta w_\phi, \quad m \in [-M, \ldots, M] \tag{5}$$

and Δw_ϕ is the sampling interval.

When a polynomial of degree K is used for the smoothed objective function at point $w_{\phi,m}$:

$$\tilde{E}(w_{\phi,m}) = \sum_{k=0}^{K} a_k m^k, \quad m \in [-M, \ldots, M] \tag{6}$$

The weight a_k is determined so that the following Mean Square Error (MSE) is minimized:

$$\varepsilon_\phi = \frac{1}{2M+1} \sum_{m=-M}^{M} (\tilde{E}(w_{\phi,m}) - E(w_{\phi,m}))^2 \tag{7}$$

As shown in [24], optimal $\mathbf{a} = [a_1, \ldots, a_M]$ is found as:

$$\mathbf{a} = (\mathbf{J}^T\mathbf{J})^{-1}\mathbf{J}^T\mathbf{w}_{s,\phi} \ , \tag{8}$$

where $\mathbf{J}$ is a Jacobian of the vector $[\tilde{E}(w_{\phi,-M}), \ldots, \tilde{E}(w_{\phi,M})]$, and its (m,k)-th element is obtained as

$$[\mathbf{J}]_{m,k} = (m-M)^k, \quad m \in [0, 2M], \ k \in [0, K] \ . \tag{9}$$

where M, Δw_ϕ and K control the smoothing rate of the objective function.

Figure 2 illustrates three iterations of the smoothing procedure to find the optimal weight for one of the features (w_{GI}). Points in Figure 2 indicate the samples taken from the objective function at each iteration, while the solid lines indicate the smoothed curves (i.e. estimated polynomials). The maximum of the smoothed curve is found and used as the starting point for the next iteration. At each subsequent iteration, the degree of smoothing is reduced by lowering Δw from 2.5×10^{-2} to 2.5×10^{-3} and then to 2.5×10^{-4}, while increasing K from 4 to 5 and 6, while keeping M constant ($M = 18$). As follows from Figure 2, the smoothing standard deviation (σ) is decreasing at each iteration of the optimization process, which indicates less smoothing and hence closer representation to the original objective function.

3.2.3 Multi-variate optimization

The multivariate optimization method to train the weights of all features with respect to the target retrieval metric is summarized in Algorithm 1. We denote the vector of feature weights by $\mathbf{w}_\phi = [w_\phi^n]_{n=1}^N$. As mentioned earlier, the weight w_ϕ^n is estimated by using $n-1$ previously estimated weights at iteration j (i.e., $\hat{w}_\phi^1, \ldots, \hat{w}_\phi^{n-1}$) and the $N-n$ estimated weights at the iteration $j-1$ (i.e., $\hat{w}_\phi^{n+1}, \ldots, \hat{w}_\phi^N$). Therefore, the univariate objective function to estimate the weight w_ϕ^n can be written as:

$$E^{n,j}(w_\phi^n) = E([\hat{w}_\phi^1, \ldots, \hat{w}_\phi^{n-1}, w_\phi^n, \hat{w}_\phi^{n+1}, \ldots, \hat{w}_\phi^N]) \tag{10}$$

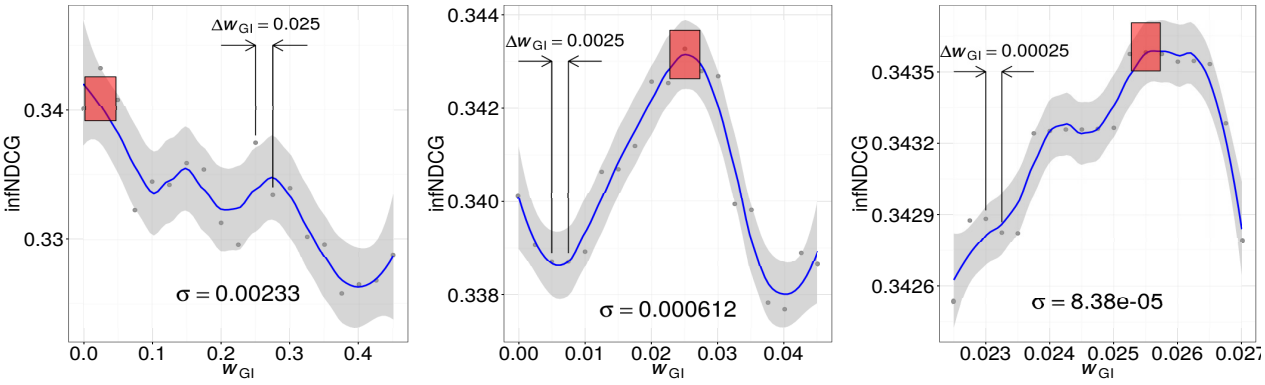
where $E^{n,j}(w_\phi^n)$ is a univariate objective function for the weight of the n-th feature at the j-th iteration.

As can be seen from Algorithm 1, first explicit and latent concepts of training queries are extracted from different sources (line 1) and then $\mathbf{w}_\phi$ is randomly initialized (line 2). At each iteration of the proposed optimization method (line 3), $\mathbf{w}_\phi$ is randomly shuffled (line 4). After that for each element of $\mathbf{w}_\phi$ (line 5) and for each sampling policy (line 6), the objective function (i.e., $E^{n,j}(w_\phi^n)$) is sampled at the points $\mathbf{w}_{s,\phi}^n = [w_{\phi,m}^n]_{m=-M}^M$ (line 7). The sampling policy determines the values of M, K, and Δw at each iteration of the optimization approach. The smoothed objective function $\tilde{E}^{n,j}(w_{\phi,m}^n)$ is obtained using the samples from $E^{n,j}(w_\phi^n)$ (line 7). Then, the optimum point of $\tilde{E}^{n,j}(w_{\phi,m}^n)$ (i.e., $\hat{w}_{\phi,m}^n$) is estimated (line 9). Next, the n-th element of $\mathbf{w}_\phi$ is replaced by its estimated value (i.e., $\hat{w}_{\phi,m}^n$) (line 10). These iterations continue until the number of iterations (i.e., j) goes beyond $j_{\max}$ (line 3) or convergence (lines 13-15).

3.3 Features

Table 1 summarizes all distinct features that are used to calculate the importance weight of each query concept c depending on its type. The list of concept types, which are determined by concept source, term representation and identification method, along with a set of features that are used to calculate the importance weight of query concepts of each type are shown in Table 2. Concepts belonging to some concept types come from only one source, while other concept types assume two sources. For example, since the concepts of type TUU are UMLS concepts that are represented by unigrams and extracted from the top retrieved documents, this concept type is associated with two concept sources (top retrieved documents and UMLS).

As can be seen from Table 2, there are four different methods for identifying explicit and latent concepts in a query. The first and simplest method is to consider all unigrams and bigrams in a query or top retrieved documents as query



(a) 1st Iteration (b) 2nd Iteration (c) 3rd Iteration

Figure 2: Application of graduated optimization to estimate the weight of the feature GI using TREC 2014 CDS track queries as the training set. Red boxes indicate the range of w_{GI} considered at the next iteration. σ is defined as the smoothing standard deviation.

Algorithm 1 Algorithm to optimize the feature weights with respect to the target retrieval metric using graduated optimization.

```
Identify explicit and latent concepts
Randomly initialize the feature weights vector (w_φ)
for j = 1 : j_max do
    Randomly shuffle w_φ
    for n = 1 : N do
        for each sampling policy do
            Sample E^{n,j}(w_φ^n)
            Obtain Ẽ^{n,j}(w_{φ,m}^n)
            Obtain the optimum point ŵ_φ^n
            Update n-th element of w_φ by ŵ_φ^n
        end for
    end for
    if Convergence then
        Break
    end if
end for
```

concepts. The second approach uses MetaMap [1] to identify UMLS concepts in a query or top-retrieved documents. The third approach uses the Wikipedia-based health relatedness measure defined in [29] as:

$$hrm(c) = \frac{P(p \text{ is health-related}|c \in p)}{1 - P(p \text{ is health-related}|c \in p)} \quad (11)$$

where $P(p \text{ is health-related}|c \in p)$ is the probability that a Wikipedia page p is health-related given that c occurs in p. Concepts for which this probability exceeds a pre-defined threshold are assumed to be health-related. The fourth approach uses the UMLS relationships table (MRREL.RRF table[2], which we further also refer to as the UMLS concept graph) to select the concepts related to the UMLS concepts identified in a query as latent concepts.

All features in Table 1, except Semantic Direction (SD), Semantic Popularity (SP) and Type Effectiveness (TE), are relatively simple and do not require a detailed explanation. Semantic direction is defined as follows. If S_c is the semantic type of concept c, S_o is the semantic type of the query concept o, to which concept c is related and $d(S_r, S)$ is the distance (i.e. the number of edges) from the root node (S_r) to node S in the UMLS semantic network, then the expansion concept c is defined to have an inward direction relative to the original concept o in the UMLS semantic network (i.e. the expansion concept is more general than the original query concept), if $d(S_r, S_c) < d(S_r, S_o)$. This feature is defined only for the UMLS expansion concepts that are related to the UMLS concepts in the original query.

Semantic popularity of concept c is defined as the number of concepts that are related to concept c in the UMLS concept graph (it can also be viewed as a node degree of concept c in the UMLS concept graph). A large value of this feature indicates popularity and generality of concept c. Type effectiveness is a binary feature that indicates whether the UMLS semantic type of concept c is effective for query expansion. As defined earlier, a semantic type is effective if its corresponding concepts can increase the precision of retrieval results when added to a query. The concept of effective semantic types for medical query expansion was first proposed in [13]. Using the training queries and relevance judgments, we fine tuned the set of effective semantic types from [13] to the collection and query sets used in this work. This will be explained in detail later.

4. EXPERIMENTS

4.1 Experimental Setup

The experimental results reported in this work were obtained using the corpus, which includes around 730,000 documents from PubMed Central (PMC), and queries from the Clinical Decision Support (CDS) track at TREC 2014 [26] and 2015 [23]. 3-fold cross-validation was used to evaluate the performance of the proposed method (`INTGR`) and the baselines, which were first trained using the query set and relevance judgments from the CDS track of TREC 2014 to maximize infNDCG, the official retrieval metric of the CDS track [26]. The proposed method and the baselines were implemented using Indri retrieval toolkit[3]. The optimal val-

[2]http://www.ncbi.nlm.nih.gov/books/NBK9685/

[3]http://www.lemurproject.org/indri/

Table 1: Brief description of features used to estimate the importance weight of concept c.

Feature	Description
TI	TF-IDF of concept c in the collection
CA	Average collection co-occurrence of concept c with other concepts in the query
CM	Maximum collection co-occurrence of concept c with other concepts in the query
NT	Number of top retrieved documents containing concept c
RS	Sum of retrieval scores of top-ranked documents containing concept c
TM	Maximum co-occurrence of concept c with other query concepts in top retrieved documents
TA	Average co-occurrence of concept c with other query concepts in top retrieved documents
GI	Do infoboxes of Wikipedia articles corresponding to concept c contain any health-related keywords?
IS	Does any of the terms of concept c exist in the title of any Wikipedia health-related articles?
CD	Average distance between concept c in the UMLS concept graph and other query, top document and related UMLS concepts identified for a query
SP	Popularity (node degree) of concept c in the UMLS concept graph
SD	Direction of concept c with respect to query concepts in the UMLS semantic network
TE	Does concept c have a UMLS semantic type that is effective for medical query expansion?

Table 2: List of types for explicit and latent query concepts along with a set of features to estimate the importance of concepts of each type (Top-docs stands for top retrieved documents for the original query).

Concept Type	Concept Source(s)	Concept Representation	Concept Extraction	Features
QU	Query	unigrams	all query unigrams	TI, NT, RS, CA, CM, TA, TM
QOB	Query	ordered bigrams	all query bigrams	TI, NT, RS, CA, CM, TA, TM
QUB	Query	unordered bigrams	all query bigrams	TI, NT, RS, CA, CM, TA, TM
QUU	Query, UMLS	unigrams	MetaMap	TI, NT, RS, CA, CM, TA, TM, TE, SP, CD
QUOB	Query, UMLS	ordered bigrams	MetaMap	TI, NT, RS, CA, CM, TA, TM, TE, SP, CD
QUUB	Query, UMLS	unordered bigrams	MetaMap	TI, NT, RS, CA, CM, TA, TM, TE, SP, CD
QDU	Query, Wikipedia	unigrams	health-relatedness measure	TI, NT, RS, CA, CM, TA, TM, GI, IS
QDOB	Query, Wikipedia	ordered bigrams	health-relatedness measure	TI, NT, RS, CA, CM, TA, TM, GI, IS
QDUB	Query, Wikipedia	unordered bigrams	health-relatedness measure	TI, NT, RS, CA, CM, TA, TM, GI, IS
TU	Top-docs	unigrams	direct identification	TI, NT, RS, CA, CM, TA, TM
TOB	Top-docs	ordered bigrams	direct identification	TI, NT, RS, CA, CM, TA, TM
TUB	Top-docs	unordered bigrams	direct identification	TI, NT, RS, CA, CM, TA, TM
TUU	Top-docs, UMLS	unigrams	MetaMap	TI, NT, RS, CA, CM, TA, TM, TE, SP, CD
TUOB	Top-docs, UMLS	ordered bigrams	MetaMap	TI, NT, RS, CA, CM, TA, TM, TE, SP, CD
TUUB	Top-docs, UMLS	unordered bigrams	MetaMap	TI, NT, RS, CA, CM, TA, TM, TE, SP, CD
TDU	Top-docs, Wikipedia	unigrams	health-relatedness measure	TI, NT, RS, CA, CM, TA, TM, GI, IS
TDOB	Top-docs, Wikipedia	ordered bigrams	health-relatedness measure	TI, NT, RS, CA, CM, TA, TM, GI, IS
TDUB	Top-docs, Wikipedia	unordered bigrams	health-relatedness measure	TI, NT, RS, CA, CM, TA, TM, GI, IS
UU	UMLS	unigrams	UMLS relationships	TI, NT, RS, CA, CM, TA, TM, TE, SP, SD, CD
UOB	UMLS	ordered bigrams	UMLS relationships	TI, NT, RS, CA, CM, TA, TM, TE, SP, SD, CD
UUB	UMLS	unordered bigrams	UMLS relationships	TI, NT, RS, CA, CM, TA, TM, TE, SP, SD, CD

ues of Dirichlet prior, Jelinek-Mercer interpolation coefficient, the sizes of ordered and unordered bigram windows in the Indri query language were empirically determined to be 2500, 0.4, 4 and 17, respectively. Figure 3 illustrates how infNDCG changes by varying the number of PRF documents (used to extract concepts) and the number of concepts extracted from PRF documents. The values of these parameters that maximize infNCDG were used in experiments using TREC 2015 CDS track queries.

Besides the proposed graduated optimization approach, we used exhaustive line search to optimize individual feature weights as another baseline (`INTGR-LS`). This method examines the parameter space in uniform increments and chooses the setting that results in the highest infNDCG. For both `INTGR` and `INTGR-LS` methods, the convergence threshold for the change in infNDCG was set to 0.001 and the number of iterations was limited to 20.

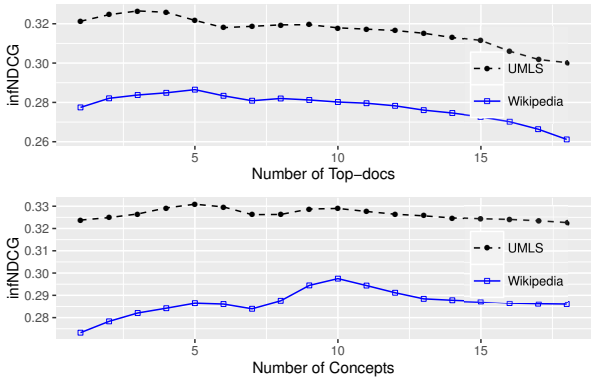


Figure 3: Average infNDCG on TREC 2014 CDS track queries by varying the number of top retrieved documents used to extract the concepts and the number of UMLS and Wikipedia concepts extracted from the top retrieved documents.

4.2 Baselines

The first baseline that was used in experiments is two-stage smoothing [36] (`Two-Stage`). Two-stage smoothing was also used as the smoothing method in implementing all other baselines and the proposed method. The other baselines used in experiments are Relevance Model (`RM`) [12], Parameterized Query Expansion (`PQE`) [5], `Wiki-Orig` and `Wiki-TD` [29], which use a Wikipedia-based health relat-

edness measure defined in (11). Other baselines that use only semantic concepts are UMLS-orig [25] and UMLS-TD [13]. UMLS-orig extracts UMLS concepts only from the query itself and breaks the phrases designating UMLS concepts into bigrams in order to incorporate them into the SDM retrieval model [17]. UMLS-TD extracts UMLS concepts from the top retrieved documents according to their semantic types. Since the original implementations of UMLS-TD and Wiki-TD are based on bag-of-words retrieval models, UMLS-TD* and Wiki-TD* are the modifications of UMLS-TD and Wiki-TD that use the SDM retrieval model to account for term dependencies when a concept is designated by a phrase.

We also compare the performance of the proposed method to the best performing methods (which used topic summaries as queries) in the CDS track of TREC in 2014 [22] and 2015 [2] (designated as TREC best). [22] used an ensemble of state-of-art unsupervised knowledge-based query expansion, re-ranking and relevance feedback methods. In [2], queries are expanded with unigrams and UMLS concepts identified in the query itself and the top retrieved documents.

4.3 Results

An initial list of 16 semantic types known to be effective for query expansion in medical records search was taken "as is" from [13]. We observed from the preliminary experiments that not all of these semantic types are effective for expansion of CDS queries. Therefore, we fine-tuned this initial list of semantic types by excluding those semantic types, for which the corresponding concepts did not improve infNDCG of retrieval results on training queries. The 5 semantic types retained from the initial list proposed in [13] are "Clinical Drug", "Disease or Syndrome", "Injury or Poisoning", "Sign or Symptom" and "Therapeutic or Preventive Procedure".

Tables 3 and 4 provide a summary of retrieval accuracy in terms of different retrieval metrics of the proposed method (INTGR) and the baselines on the query sets from the CDS track of TREC 2014 and 2015. As can be seen from Table 3, Wiki-TD* is the best performing baseline (since the best performing TREC methods are different for different query sets, they are not considered as the best performing baselines). Furthermore, the proposed algorithm outperforms INTGR-LS and the best methods in TREC 2014 and 2015.

Table 4 shows the degree of improvement and its statistical significance of the proposed method over the three best performing baselines (i.e., PQE, Wiki-TD, Wiki-TD*) and INTGR-LS. As follows from Table 4, INTGR significantly outperforms all of the best performing baselines in terms of all retrieval metrics. Using graduated non-convexity as a univariate optimization method results in 5-9% improvement of retrieval accuracy in terms of infNDCG, 10-23% improvement in terms of infAP and 8% improvement in terms of P@5 on different query sets.

Table 5 illustrates the effect of using different knowledge bases in conjunction with INTGR on its performance in terms of different evaluation metrics. As follows from Table 5, using INTGR only with Wikipedia results in the smallest improvement of retrieval accuracy across all retrieval metrics (and even a decrease of P@5). It also follows from this table that using INTGR with UMLS results in significantly greater improvement of all retrieval metrics, while the biggest improvement is achieved when explicit and latent concepts of a query are extracted from both UMLS and Wikipedia.

Figure 4 provides performance comparison of INTGR with all of the baselines in terms of P@k for k from 1 to 10 (with a step size of 1). As can be seen from this figure, for all values of k except $k = 1$ in case of TREC 2014 CDS track queries, INTGR significantly outperforms all other baselines. It also follows from Figure 4 that for most of the values of k, the methods that expand the queries with the concepts extracted from the top-ranked documents (RM, UMLS-TD, UMLS-TD*, PQE, Wiki-TD, Wiki-TD* and INTGR) outperform the methods that represent the queries with the concepts extracted from them (Wiki-Orig and UMLS-orig). The average improvements of INTGR in terms of P@k for different values of k over the weakest and strongest baselines are 0.1560 and 0.0380, respectively, on the query set from TREC 2014 CDS track, while on the query set from TREC 2015 CDS track the improvements are 0.0988 and 0.0481, respectively.

Figure 5 illustrates topic level differences between the retrieval accuracy of INTGR in terms of infNDCG with the best performing baselines (Wiki-TD* for the CDS track of TREC 2014 and PQE for the CDS track of TREC 2015) on both query sets. From Figure 5(a), it follows that infNDCG of INTGR is greater than that of Wiki-TD* on 67% of the queries in the CDS track of TREC 2014, while from Figure 5(b) it follows that infNDCG of INTGR is greater than that of PQE on 73% of the queries in the CDS track of TREC 2015. The average improvement of INTGR over Wiki-TD* in terms of infNDCG on TREC 2014 CDS track queries is 0.0518 with standard deviation 0.12, while the average improvement of INTGR over PQE in terms of infNDCG on TREC 2015 CDS track queries is 0.0345 with standard deviation 0.0734. The topics, on which INTGR has the greatest improvement and decline relative to Wiki-TD* in terms of infNDCG among those used in TREC 2014 CDS track are 16 (with 0.4593 improvement) and 14 (with 0.1462 decline). We can also observe that on the query set of TREC 2015 CDS track INTGR has the greatest improvement of 0.3026 and the greatest decline of 0.0512 in terms of infNDCG on topics 6 and 8, respectively. Figure 6 also provides a detailed comparison of retrieval accuracy of INTGR in terms of infNDCG with the best performing baselines (Wiki-TD* for TREC 2014 CDS track and PQE for TREC 2015 CDS track) at the level of each individual topic in the CDS track of TREC 2014 and 2015.

We continued empirical evaluation of INTGR by analysis of its performance on difficult queries. We define a query as *difficult* if infNDCG of Two-Stage on this query is less than 0.1 and as *very difficult* if infNDCG of Two-Stage is less than 0.05. We observed that INTGR outperformed Wiki-TD* on 59% of difficult queries and on 86% of very difficult queries in the CDS track of TREC 2014. We also observed that INTGR outperformed PQE on 56% of difficult queries and on 77% of very difficult queries in CDS track of TREC 2014.

4.4 Discussion

Based on experimental analysis of INTGR presented in the previous section, we can conclude that the subset of UMLS semantic types that are effective for expansion of CDS queries is fairly small (includes less than 4% of UMLS semantic types). These semantic types can be grouped into three categories: "Disorders", "Chemical & Drugs" and "Procedures". These three categories in turn can be conceptually mapped

Table 3: Summary of retrieval accuracy of the proposed method and the baselines on the query sets from the CDS track of TREC 2014 and 2015.

Query set	TREC 2014 CDS track			TREC 2015 CDS track		
Method	infNDCG	infAP	P@5	infNDCG	infAP	P@5
Two-Stage [36]	0.1945	0.0493	0.3533	0.2110	0.0449	0.4200
Wiki-Orig [29]	0.2069	0.0550	0.3533	0.2193	0.0457	0.4133
UMLS-Orig [25]	0.2074	0.0569	0.3867	0.2206	0.0478	0.4400
RM [12]	0.2662	0.0836	0.4400	0.2765	0.0740	0.4600
UMLS-TD [13]	0.2577	0.1523	0.4067	0.2429	0.0748	0.4600
UMLS-TD*	0.2724	0.0810	0.4133	0.2503	0.0614	0.4667
PQE [5]	0.2796	0.0873	0.4733	0.2792	0.0762	0.4400
Wiki-TD [29]	0.2764	0.0881	0.4467	0.2418	0.0597	0.4267
Wiki-TD*	0.2883	0.0944	0.4600	0.2519	0.0633	0.4600
TREC best [22, 2]	0.2631	0.0757	0.4067	0.2928	0.0777	0.4467
INTGR-LS	0.3114	0.0993	0.4867	0.2987	0.0792	0.4800
INTGR	**0.3401**	**0.1229**	**0.5267**	**0.3135**	**0.0873**	**0.5200**

Table 4: Statistical significance and improvement in retrieval accuracy of the proposed method (INTGR) relative to its modification (INTGR-LS) and three best performing baselines (Wiki-TD, PQE and Wiki-TD*) on the query sets from the CDS track of TREC 2014 and 2015. $\star$ and † indicate statistically significant improvement with $p < 0.05$ and $p < 0.1$, respectively.

Query set	TREC 2014 CDS track			TREC 2015 CDS track		
Method	infNDCG	infAP	P@5	infNDCG	infAP	P@5
Wiki-TD	23.05%$\star$†	39.50%$\star$†	17.91%$\star$†	29.65%$\star$†	46.23%$\star$†	23.81%†
PQE	21.64%$\star$†	40.78%$\star$†	11.28%†	12.28%†	14.56%$\star$	18.18%$\star$†
Wiki-TD*	17.97%$\star$†	30.19%$\star$†	14.50%$\star$†	24.45%$\star$†	37.91%$\star$†	13.04%$\star$†
INTGR-LS	9.22%$\star$†	23.77%$\star$†	8.22%$\star$†	4.95%$\star$†	10.22%$\star$	8.33%$\star$†

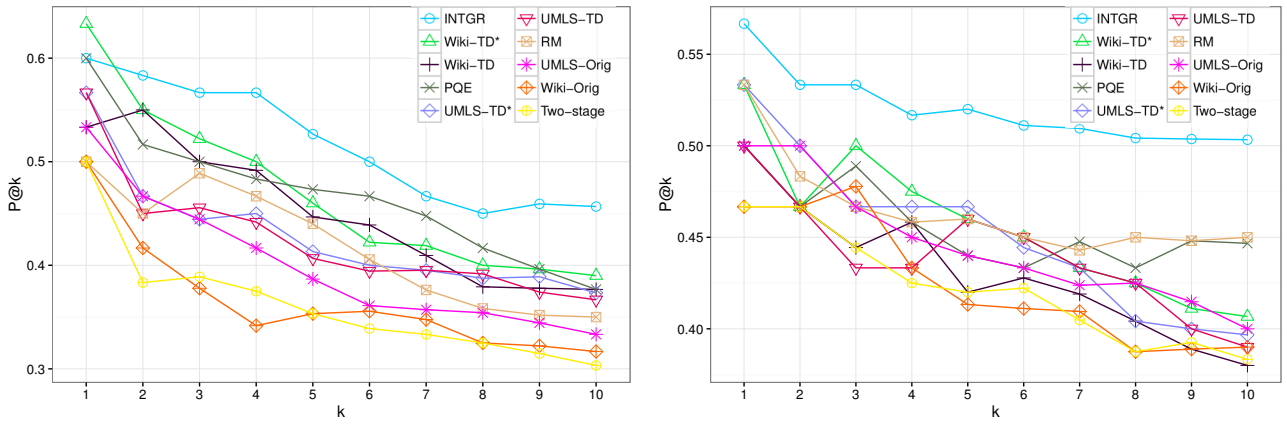
Table 5: Comparison of effectiveness of different knowledge bases on the query sets from the CDS track of TREC 2014 and 2015.

Query set	TREC 2014 CDS track			TREC 2015 CDS track		
Method	infNDCG	infAP	P@5	infNDCG	infAP	P@5
INTGR using no knowledge bases	0.2673	0.0875	0.4601	0.2771	0.0758	0.4633
INTGR using only Wikipedia	0.2975 (11.30%)	0.0936 (6.97%)	0.4533 (-1.47%)	0.2954 (6.60%)	0.0779 (2.77%)	0.4667 (0.09%)
INTGR using only UMLS	0.3309 (23.79%)	0.1170 (33.71%)	0.5200 (13.02%)	0.3012 (8.67%)	0.0786 (3.93%)	0.5033 (7.93%)
INTGR using UMLS and Wikipedia	0.3401 (27.23%)	0.1229 (40.46%)	0.5267 (14.47%)	0.3135 (13.14%)	0.0873 (15.17%)	0.5200 (11.52%)

to the three main types of CDS queries: "Diagnosis", "Treatment" and "Test".

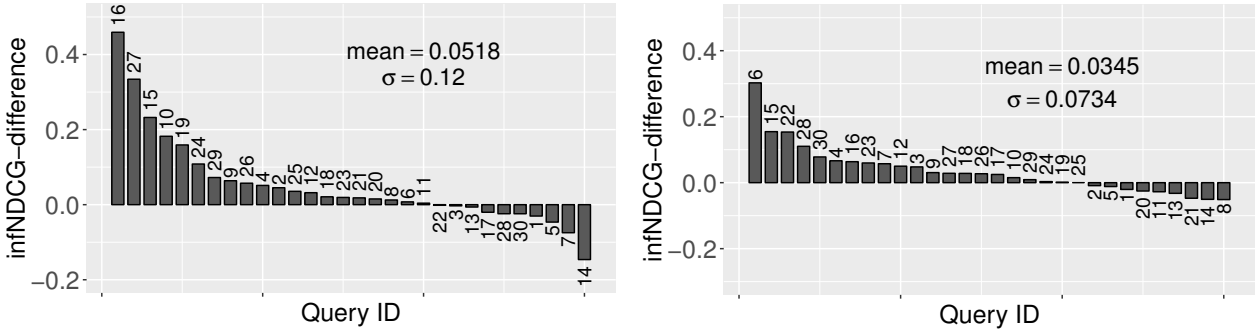
From tables 3 and 4, it follows that the proposed query representation method significantly outperforms all baselines in terms of all evaluation metrics and on both training and evaluation query sets. Furthermore, although INTGR was trained on the CDS track queries of TREC 2014 with the goal of maximizing infNDCG, INTGR also achieved significant (and, in many cases, even greater) improvement over the baselines in terms of other evaluation metrics (i.e., infAP and P@5) on both training and testing query sets. Also, as can be seen from Tables 3 and 4, the proposed method has significantly better performance when it is used in conjunction with graduated optimization method (INTGR) than when it is used with exhaustive line search (INTGR-LS), which we attribute to the ability of graduated optimization to efficiently find global optima of non-smooth and non-convex objective functions. Line search, on the other hand, may miss global optima, if the step size is not sufficiently small. In general, choosing the appropriate step-size is non-trivial and can dramatically affect the performance of line search.

As follows from Table 3, methods that utilize semantic (Wiki-TD/Wiki-TD* and UMLS-TD/UMLS-TD*) and statistical (RM and PQE) concepts for query representation and expansion behave differently on training and evaluation query sets. In particular, methods using semantic concepts show better results than the methods based on statistical concepts on the training query set, while the methods based on statistical concepts show better results on evaluation query set. However, the proposed method (INTGR) provides excellent results on both query sets, which indicates the utility of accounting for both types of concepts in a retrieval method for CDS queries. On the other hand, Table 5 demonstrates that for the methods based on semantic concepts, UMLS is a better choice than Wikipedia with respect to all metrics, if only



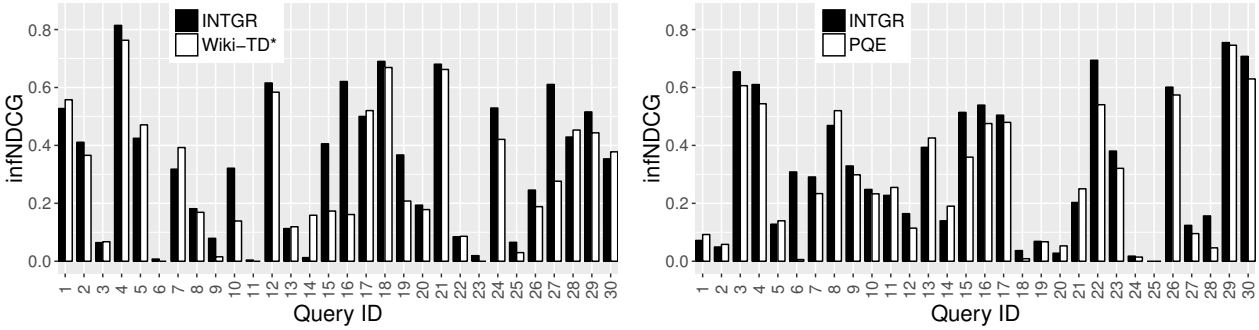
(a) TREC 2014 CDS track topics (b) TREC 2015 CDS track topics

Figure 4: Comparison of INTGR with the baselines in terms of P@k for $k \leq 10$ on the query sets from the CDS track of TREC 2014 and 2015.



(a) TREC 2014 CDS track topics (b) TREC 2015 CDS track topics

Figure 5: Topic-level differences of the infNDCG values for INTGR and the best-performing baselines (Wiki-TD* for TREC 2014 CDS track and PQE for TREC 2015 CDS track).



(a) TREC 2014 CDS track (b) TREC 2015 CDS track

Figure 6: Topic-level comparison of the infNDCG values for INTGR, the best performing baselines (Wiki-TD* for TREC 2014 CDS track and PQE for TREC 2015 CDS track).

one knowledge repository is used. However, as follows from Table 5, combining both knowledge bases results in better retrieval accuracy than using any one of them individually.

Although from Figures 4 and 5 as well as Tables 3 and 4 it follows that `INTGR` has slightly lower accuracy improvement over its best-performing baseline and `Two-Stage` on the testing query set than on the training query set, the improvement that `INTGR` achieves over `Two-Stage` is much higher than the improvement of the best performing baseline over `Two-Stage`. However, as follows from Figures 6 and

5, there is a greater number of topics on which INTGR has better retrieval accuracy than the best performing baseline on both training and testing query sets. Therefore, based on these observations, we can conclude that INTGR is robust to overfitting, due to its use of multiple and diverse relevance signals and concept sources.

5. CONCLUSION

In this paper, we proposed a method to represent CDS queries using statistical and semantic concepts from the query, top retrieved documents and knowledge bases. Our work logically extends previous research, which focused only on studying the utility of statistical query concepts [4], semantic query concepts [3], statistical and semantic query concepts [7], statistical [17, 5] and semantic [29] concepts from the query and top retrieved documents for query expansion. Experiments using a collection of PubMed articles and TREC Clinical Decision Support (CDS) track queries indicate that the proposed method significantly outperforms state-of-the-art baselines for ad hoc and medical IR.

Acknowledgments

This work was partially supported by the National Institutes of Health under the grant R21 DK108071-01A1.

6. REFERENCES

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

[2] S. Balaneshin-kordan, A. Kotov, and R. Xisto. WSU-IR at TREC 2015 clinical decision support track: Joint weighting of explicit and latent medical query concepts from diverse sources. In *Proceedings of TREC'15*, 2015.

[3] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *Proceedings of SIGIR'08*, pages 491–498, 2008.

[4] M. Bendersky, D. Metzler, and W. B. Croft. Learning concept importance using a weighted dependence model. In *Proceedings of WSDM'10*, pages 31–40, 2010.

[5] M. Bendersky, D. Metzler, and W. B. Croft. Parameterized concept weighting in verbose queries. In *Proceedings of SIGIR'14*, pages 605–614, 2011.

[6] A. Blake and A. Zisserman. *Visual reconstruction*, volume 2. MIT press Cambridge, 1987.

[7] S. Choi, J. Choi, S. Yoo, H. Kim, and Y. Lee. Semantic concept-enriched dependence model for medical information retrieval. *Journal of Biomedical Informatics*, 47:18–27, 2014.

[8] J. I. Garcia-Gathright, F. Meng, and W. Hsu. UCLA at TREC 2014 clinical decision support track: Exploring language models, query expansion, and boosting. *Proceedings of TREC'14*, 2014.

[9] B. Koopman, G. Zuccon, P. Bruza, L. Sitbon, and M. Lawley. Information retrieval as semantic inference: a graph inference model applied to medical search. *Information Retrieval Journal*, 19(1-2):6–37, 2016.

[10] B. Koopman, G. Zuccon, A. Nguyen, D. Vickers, L. Butt, and P. Bruza. Exploiting SNOMED CT concepts & relationships for clinical information retrieval: Australian e-Health Research Centre and Queensland University of Technology at the TREC 2012 medical track. *Proceedings of TREC'12*, 2012.

[11] A. Kotov and C. Zhai. Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In *Proceedings of WSDM'12*, pages 403–412, 2012.

[12] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of SIGIR'01*, pages 120–127, 2001.

[13] N. Limsopatham, C. Macdonald, and I. Ounis. Inferring conceptual relationships to improve medical records search. In *Proceedings of OAIR'13*, pages 1–8, 2013.

[14] J. Lin and D. Demner-Fushman. The role of knowledge in conceptual retrieval: a study in the domain of clinical medicine. In *Proceedings of SIGIR'06*, pages 99–106, 2006.

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

[16] D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proceedings of SIGIR'05*, pages 472–479, 2005.

[17] D. Metzler and W. B. Croft. Latent concept expansion using Markov random fields. In *Proceedings of SIGIR'07*, pages 311–318, 2007.

[18] D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. *Information Retrieval*, 10(3):257–274, 2007.

[19] D. A. Metzler, W. B. Croft, and A. McCallum. Direct maximization of rank-based metrics for information retrieval. Technical report, CIIR, 2005.

[20] H. Mobahi and J. W. Fisher III. On the link between gaussian homotopy continuation and convex envelopes. In *EMMCVPR'15*, pages 43–56, 2015.

[21] W. Morgan, W. Greiff, and J. Henderson. Direct maximization of average precision by hill-climbing, with a comparison to a maximum entropy approach. In *Proceedings of NAACL-HLT'04*, pages 93–96, 2004.

[22] A. Mourao, F. Martins, and J. Magalhaes. NovaSearch at TREC 2014 clinical decision support track. *Proceedings of TREC'14*, 2014.

[23] K. Roberts, M. S. Simpson, E. Voorhees, and W. R. Hersh. Overview of the trec 2015 clinical decision support track. *Proceedings of TREC'15*, 2015.

[24] R. W. Schafer. What is a Savitzky-Golay filter?[lecture notes]. *IEEE Signal Processing Magazine*, 28(4):111–117, 2011.

[25] W. Shen, J.-Y. Nie, X. Liu, and X. Liui. An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM@ CLEF2014eHealthTask 3. *Proceedings of the ShARe/CLEF eHealth Evaluation Lab*, 2014.

[26] M. S. Simpson, E. M. Voorhees, and W. Hersh. Overview of the trec 2014 clinical decision support track. *Proceedings of TREC'14*, 2014.

[27] C. A. Sneiderman, D. Demner-Fushman, M. Fiszman, N. C. Ide, and T. C. Rindflesch. Knowledge-based methods to help clinicians find answers in MEDLINE. *Journal of the American Medical Informatics Association*, 14(6):772–780, 2007.

[28] L. Soldaini, A. Cohan, A. Yates, N. Goharian, and O. Frieder. Query reformulation for clinical decision support search. *Proceedings of TREC'14*, 2014.

[29] L. Soldaini, A. Cohan, A. Yates, N. Goharian, and O. Frieder. Retrieving medical literature for clinical decision support. In *Advances in Information Retrieval*, pages 538–549. Springer, 2015.

[30] P. Sondhi, J. Sun, C. Zhai, R. Sorrentino, and M. S. Kohn. Leveraging medical thesauri and physician feedback for improving medical literature retrieval for case queries. *Journal of the American Medical Informatics Association*, 19(5):851–858, 2012.

[31] P. Srinivasan. Retrieval feedback in MEDLINE. *Journal of the American Medical Informatics Association*, 3:157–167, 1996.

[32] C. Wang and R. Akella. Concept-based relevance models for medical and semantic information retrieval. In *Proceedings of CIKM'15*, pages 173–182, 2015.

[33] Y. Wang and H. Fang. Exploring the query expansion methods for concept based representation. *Proceedings of TREC'14*, 2014.

[34] Z. Xie, Y. Xia, and Q. Zhou. Incorporating semantic knowledge with MRF term dependency model in medical document retrieval. In *NLPCC'15*, pages 219–228. Springer, 2015.

[35] Y. Xu, G. J. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of SIGIR'02*, pages 59–66, 2009.

[36] C. Zhai and J. Lafferty. Two-stage language models for information retrieval. In *Proceedings of SIGIR'02*, pages 49–56, 2002.

[37] M. Zhong and X. Huang. Concept-based biomedical text retrieval. In *Proceedings of SIGIR'06*, pages 723–724, 2006.

[38] W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *Proceedings of SIGIR'07*, pages 655–662, 2007.