Learning to Handle Negated Language in Medical Records Search

Nut Limsopatham[1], Craig Macdonald[2], Iadh Ounis[2]
nutli@dcs.gla.ac.uk[1], firstname.lastname@glasgow.ac.uk[2]

School of Computing Science
University of Glasgow, Glasgow, UK

ABSTRACT

Negated language is frequently used by medical practitioners to indicate that a patient does not have a given medical condition. Traditionally, information retrieval systems do not distinguish between the positive and negative contexts of terms when indexing documents. For example, when searching for patients with angina, a retrieval system might wrongly consider a patient with a medical record stating "no evidence of angina" to be relevant. While it is possible to enhance a retrieval system by taking into account the context of terms within the indexing representation of a document, some non-relevant medical records can still be ranked highly, if they include some of the query terms with the intended context. In this paper, we propose a novel learning framework that effectively handles negated language. Based on features related to the positive and negative contexts of a term, the framework learns how to appropriately weight the occurrences of the opposite context of any query term, thus preventing documents that may not be relevant from being retrieved. We thoroughly evaluate our proposed framework using the TREC 2011 and 2012 Medical Records track test collections. Our results show significant improvements over existing strong baselines. In addition, in combination with a traditional query expansion and a conceptual representation approach, our proposed framework could achieve a retrieval effectiveness comparable to the performance of the best TREC 2011 and 2012 systems, while not addressing other challenges in medical records search, such as the exploitation of semantic relationships between medical terms.

Categories and Subject Descriptors: H.3.3 [Information Search & Retrieval]: Search process

Keywords: Medical Records Search; Negation; Regression-trees

1. INTRODUCTION

Electronic medical records (EMRs) have been used to document the medical history of patients to ensure patient safety and prevent medical errors [30, 32]. These EMRs can be used to improve the quality of healthcare services.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2505706.

Table 1: Examples of sentences in EMRs where the presence of the query term 'cancer' does not always indicate the relevance.

Patient admitted with **cancer**
Diagnosed and found no evidence of **cancer**
Negative result on **cancer** screening test

Moreover, EMRs can also be used to evaluate the effectiveness of a particular medical procedure. For example, medical researchers may search from the EMRs for the patients with particular medical condition(s), when forming a clinical trial for a particular treatment [14, 35, 36]. This requires an effective information retrieval (IR) system that is able to cope with the special characteristics of medical records and queries. To facilitate research on searching medical records, TREC initiated the Medical Records track in 2011 [35, 36].

One of the major challenges of searching from medical records is the use of negated language [18, 20]. Negation is commonly used in medical records to indicate that the patient does not suffer from a particular medical condition [16]. As a result, the presence of a query term does not always imply that the record is relevant to the query [4]. In particular, the relevance also depends on the context of query terms occurring in the medical records. For example, while all the three sentences shown in Table 1 contain the query term 'cancer', only the first sentence indicates that the disease is pertaining to the patient, while the other two sentences are irrelevant towards the query searching for patients suffering from 'cancer'. Averbuch et al. [4] estimated that ignoring negation in medical records could result in a drop of 40% in retrieval performance. In addition, several top performing search systems (e.g. [11, 17, 20, 24, 39]) at the TREC 2011 and 2012 Medical Records track showed that dealing with negation in medical records improved the retrieval performance. King et al. [17] reported that by removing the negated sentences from a search system, the retrieval performance could be improved by 5%. In particular, the approaches described in [11, 17, 39] disregard all parts of sentences having a negative context during the search process. Different from the others, Limsopatham et al. [20, 24] proposed a representation approach that could significantly improve the retrieval effectiveness, by representing terms having either negative or positive contexts differently. However, their representation approach [20] could only help to retrieve medical records containing query terms with the correct contexts, but it does not necessarily prevent medical records containing query terms with opposite contexts from being retrieved. For example, they represent a medical record such as "heart disease patient with no history of diabetes" as *"heart disease n$diabetes"*, where n$diabetes is the

Table 2: Comparison of approaches that handle negated language in medical records search.

Approach	Improve query representation	Discarding negated language	Improve document representation	Learning to penalise non-relevant documents
Traditional approach (e.g. [18])	✗	✗	✗	✗
Post-retrieval filtering [29, 37]	✔	✗	✗	✗
Vector negation [37]	✔	✗	✗	✗
Boolean retrieval model	✔	✗	✗	✗
Discarding negated sentences (e.g. [11, 17, 39])	✗	✔	✔	✗
Representing the context of terms [20]	✗	✗	✔	✗
Our proposed framework to handle negation	✔	✗	✔	✔

negative version of diabetes. In this case, for a typical best match retrieval model, a query such as "find patients with heart disease and diabetes" would still retrieve this medical record, since two of the three non-stopped terms (i.e. heart, disease) are matched, even though it is a medical record of a patient who is known not to have diabetes.

In this paper, we propose a novel learning framework to prevent medical records clearly stating that their associated patients do not have the medical conditions stated in the query from being ranked highly. In particular, our framework consists of three components. Firstly, as we intend to promote medical records having query terms with their intended context and to demote those containing the query terms with the opposite context, we follow [20] and represent terms in medical records and queries by taking their contexts into account. Secondly, we penalise the medical records containing the query terms with the opposite context, which we refer to as the *opposite context terms*, in order to prevent medical records having the occurrences of the query terms with the opposite intended context from being ranked highly. For example, for a query 'hypertension', the opposite context term is 'n$hypertension'. Finally, we set an effective penalising weight for each of the opposite context terms, to reduce the relevance score of medical records containing these terms. Specifically, we deploy a regression technique to identify the penalising weight of an opposite context term using features (e.g. term frequency, and co-occurrence information), obtained from the query and the medical records.

We evaluate our proposed framework using the standard test collections provided by the TREC 2011 [36] and 2012 [35] Medical Records track. Our results show that the proposed framework can significantly outperform a strong negation handling baseline. When combined with a conceptual representation using an existing approach [31], our achieved retrieval performance is comparable to those of the best TREC 2011 and 2012 systems.

The contributions of this paper are three-fold:

- We introduce a novel approach to handle negation in medical records search by penalising medical records containing the opposite context terms and preventing them from being ranked highly, since they may be irrelevant to the query (as indicated by the occurrence of the opposite context terms).
- We propose a novel supervised learning framework to estimate the weight of each opposite context term using a number of features (e.g. term frequency, co-occurrence between terms with positive and negative contexts) to ensure that medical records containing the opposite context terms are effectively penalised.
- We thoroughly evaluate our proposed framework with a medical records search system using the standard experimentation setup provided by the TREC 2011 and 2012 Medical Records track.

The remainder of the paper is organised as follows. In Section 2, we discuss related work and position our paper in the literature. Section 3 introduces our proposed learning framework. Section 4 discusses our proposed learning procedure. Section 5 describes the experimental setup. We empirically evaluate our framework in Section 6.1. In Section 7, we further evaluate our framework when combined with a conceptual representation approach, which has been shown to be effective for this medical records search task [21]. Finally, we conclude the paper and discuss future directions in Section 8.

2. RELATED WORK

The wide-spread use of negated language is a major challenge in the searching of medical records. Indeed, negated language is extensively used in medical records by practitioners to indicate that the patients do not possess a particular symptom or condition [18, 20]. For example, a medical record stating that "the patient has a cough but denies fever" indicates that the patient has a cough but does not have fever, which is not relevant to a query issued to find patients having both cough and fever. This causes a problem for IR systems that estimate relevance from only the matching between the terms in a document and a query. However, prior work dealing with negation in documents is limited. Table 2 summarises and compares our work in this paper with related work. Specifically, classical/traditional IR approaches (e.g. BM25 [28]) do not explicitly deal with negation. They simply ignore the presence of negated language in medical records and query [18]. On the other hand, most of the previous works that deal with negated language (e.g. boolean retrieval model, post-retrieval filtering [29, 37], vector negation [37]) only tackled it within queries. For example, the boolean retrieval model focuses on retrieving documents by firstly forming a list of documents containing the query terms and then removing the documents with the occurrence of the negated query terms from the retrieved list. Nevertheless, these approaches do not take into account negated language in documents. Indeed, most phrases indicating negation (e.g. no, not) are commonly seen as stopwords and are usually discarded during indexing [10]. In contrast, there have been recent attempts to deal with negation in medical records search (e.g. [11, 17, 20, 24, 39]). In particular, these approaches commonly deploy a negation detection tool (e.g. NegEx [10] or NegFinder [26]) to deal with negated language in medical records. For example,

King et al. [17] and Zhu et al. [39] proposed to effectively disregard terms with the negated context from a medical search system. Specifically, they removed from a search system parts of the sentences that contain the negated context identified using the NegEx tool. Limsopatham et al. [20] introduced a document representation approach that distinguishes the negated terms from their corresponding terms with the positive context within a search system. Indeed, they firstly deployed the NegEx tool to identify negated terms from the medical records. Then, during indexing, they represent a term and its corresponding term with the opposite context differently. Hence, during retrieval, medical records do not gain a relevance score from negated query terms, if the query terms are in a positive context.

However, all aforementioned approaches suffer from a number of limitations. Indeed, traditional approaches that simply ignore the presence of negated language in medical records and queries might prevent a search system from being effective, since the context of the terms in the search system is not recognised. On the other hand, while approaches such as post-retrieval filtering [29, 37] and vector negation [37] improve the representation of the queries, they are still ineffective in demoting the medical records containing the opposite context terms, because they do not alter the representation of the medical records. Conversely, the approaches that alter the medical records representation in order to discard the negated sentences from the search system (e.g. [11, 17, 39]) cannot cope with the negated language in the queries. For example, for a query searching for "heart disease patients who have no history of diabetes", the approaches in [11, 17, 20, 39] might retrieve the medical records of patients suffering from both heart disease and diabetes, since the terms 'heart' and 'disease' are still matched with the query, even if these medical records are not relevant.

To overcome these limitations, we propose a novel learning framework to effectively handle negation in medical records search systems. Specifically, we propose to improve the representation of both medical records and queries by representing terms along with their context. Moreover, we penalise medical records containing the opposite context terms to prevent these medical records from being ranked highly, when they are unlikely to be relevant.

Machine learning techniques, such as linear and logistic regressions, have been used to identify weights for query terms when scoring documents. For instance, Cao et al. [9] used SVM for term classification. They distinguished between good and bad terms for query expansion and expanded a query with only those deemed good expanded terms by taking the classification score into account. Lease et al. [19] introduced *Regression Ranking*, which deploys a linear regression to estimate term weights from the past queries using features, such as term and document frequencies and part-of-speech. Later, Bendersky et al. [7] proposed a linear regression approach to parameterise the weights of the terms. As regression has been shown to be effective for estimating term weights, in this work, we also deploy regression to estimate the weight of the opposite context terms, to ensure that the relevance score of the medical records with occurrences of these terms are properly penalised.

3. A LEARNING FRAMEWORK TO HANDLE NEGATION

In this section, we describe our new *Learning framework To Handle Negation (LTHN)* in medical records search. The proposed framework addresses the drawbacks of the existing negation handling approaches, discussed in Section 2. The major difference between our framework and those existing approaches in the literature is that our framework goes beyond matching terms with the correct context between a medical record and a query, by penalising the medical records that contain the opposite context terms. This prevents these medical records from being retrieved, since they are likely to be non-relevant. Specifically, our framework consists of three components:

1. *Context identification*, to identify and represent the context of the terms in both medical records and queries;
2. *Context-based penalisation*, to model the penalisation of medical records containing the opposite context terms when ranking medical records;
3. *Penalising weight estimation*, to accurately weight the opposite context terms to prevent medical records that may be irrelevant to the query from being ranked highly.

Next, we discuss in detail the context identification component in Section 3.1. In Section 3.2, we describe the context-based penalisation component to demote non-relevant medical records. Finally, Section 3.3 introduces the penalising weight estimation component that deploys a regression technique to estimate the weight of the opposite context terms for penalising medical records containing these terms.

Table 3: An example of a sentence processed using the context identification component – italicised terms have negative context.

Original sentence	patient with lung cancer who does not have diabetes
Identified negative context	patient with lung cancer who does not *have diabetes*
Removing stopwords	lung cancer *diabetes*
Context Identification	lung cancer *n$diabetes*

3.1 Context Identification

The context identification phase is an important component of our framework, as it helps a search system to distinguish between a term with different contexts (e.g. diabetes and no diabetes). In particular, this component preprocesses medical records and queries by using a negation detection tool to identify negated terms. In this work, we follow [20] and use the NegEx algorithm [10] to differentiate between terms having positive and negative contexts in each sentence in both the medical records and queries. Then, terms with the negative context are replaced with their negative version before processing in an IR system. Table 3 shows an example of how our context identification component deals with a sentence, such that a term 'diabetes' which has a negative context is replaced with its negated form, 'n$diabetes'. This allows the IR system to match both terms and their context during retrieval.

However, even though the context identification component can improve the representation of medical records and queries, it could not prevent non-relevant medical records that contain some of the query terms with the correct context from being retrieved. We introduce the second component to deal with this problem in the next section.

3.2 Context-based Penalisation

To decrease the likelihood that non-relevant medical records (indicated by the occurrence of the opposite context terms) are retrieved, the second component of our framework penalises these medical records based on the occurrences of the

Table 4: An example of how our framework deals with negation, where w_n is the weight of a term.

Context identified EMR	lung cancer n$diabetes
Context identified query	diabetes lung cancer
Context-based penalisation	diabetes lung cancer *n$diabetic*$\cdot w_1$ *n$lung*$\cdot w_2$ *n$cancer*$\cdot w_3$

opposite context terms. In particular, if a query searches for a particular context (e.g. positive) of a term but a medical record contains the query term with the opposite context (e.g. negative), the medical record is likely to be non-relevant. For example, for a query "find patient with diabetes and lung cancer", a medical record stating "patient with lung cancer who does not have diabetes" is non-relevant, since the medical record clearly states that the associated patient does not have a medical condition that the query is searching for (i.e. diabetes). However, as shown in Table 4, with only the context identification component, the record is represented as "lung cancer n$diabetes". As a result, the medical record may still be retrieved since it matches two of the three query terms (i.e. 'lung' and 'cancer'). The context-based penalisation component copes with this issue by reducing the relevance score of medical records, if they contain a term t' with the opposite context to its corresponding query term t (e.g. 'n$diabetes' is the opposite context term corresponding to query term 'diabetes'). This component models the relevance score of a medical record based on both the occurrence of the query terms and the opposite context terms, so that the relevance score of the medical records containing the opposite context terms will be penalised, while the relevance score will be increased if the query terms with the correct context occur in the medical records. To do so, the terms having the opposite context to their corresponding query terms are added to the query with a particular weight to penalise the relevance score of medical records containing these opposite context terms. For example, in Table 4, 'n$diabetes', 'n$lung' and 'n$cancer', which are the opposite context terms of the query terms 'diabetes', 'lung' and 'cancer', respectively, are added to the query with the penalising weight w_n. Equation (1) shows how the second component of the framework calculates the relevance score of a medical record d towards a query Q.

$$score_{context\_penalise}(d,Q) = \sum_{t \in Q} score(d,t) + \sum_{t' \in opposites(Q)} w(t') \cdot score(d,t') \quad (1)$$

where $opposites(Q)$ returns a set of the opposite context terms (e.g. 'n$diabetes', 'n$lung' and 'n$cancer' are the opposite context terms of the query illustrated in Table 4), $w(t')$ is the weight of an opposite context term t' (e.g. w_1, w_2 and w_3 in Table 4) – typically $w(t') < 0$, to penalise the occurrences of t'. $score(\cdot)$ can be calculated using any term weighting model, such as BM25 [28]. Indeed, the first part of the equation is the classical document scoring approach that estimates the relevance of a medical record based on the appearance of a query term t. On the other hand, the second part of Equation (1) aims to penalise the medical records that contain an opposite context term t' (e.g. 'n$diabetes').

From Equation (1), we draw attention to $w(t')$, which is a crucial parameter for effectively penalising a medical record for the occurrences of t'. Indeed, there are different alternatives to estimate $w(t')$, such as giving a fixed weight to all the opposite context terms. However, to effectively estimate the weight of the opposite context terms, in the next section, we introduce the last component of our framework, which deploys a regression technique to learn an effective weight for the opposite context terms using several statistical features.

3.3 Penalising Weight Estimation

The penalising weight estimation component focuses on finding an appropriate weight to penalise the medical records containing an opposite context term (i.e. $w(t')$ in Equation (1)), to prevent these medical records from being ranked highly, hence leading to an effective retrieval performance. We hypothesise that not all opposite context terms are equally important for penalising the relevance of medical records. For example, consider the query "find a patient with diabetes and lung cancer" in Table 4, which is represented as "diabetes lung cancer". For this query, medical records with the term 'n$diabetes' should be penalised more than those containing 'n$cancer', as in the former, it is likely that the patient does not suffer from diabetes, while a medical record containing 'n$cancer' may discuss a patient who does not have another type of cancer (e.g. the patient does not have kidney cancer). In this way, the discriminative power of the opposite context terms should be considered when assigning the penalising weights.

We view this problem of estimating the penalising weights of different opposite context terms as a supervised learning problem, where the objective is to predict an estimated effective penalising weight for each opposite context term, based on the retrieval performance on a training set. By doing so, we benefit from the fact that several features (e.g. term frequency and co-occurrence statistics) of the opposite context terms are taken into account to estimate the penalising weights. Indeed, a regression function $f(\cdot)$ calculates the penalising weight of an opposite context term t' using a set of features $\Phi^{t'}$, which are associated to the term t', as follows:

$$w(t') = f(\Phi^{t'}) \quad (2)$$

The regression function $f(\Phi^{t'})$ aims to approximate the weight $w(t')$ using a particular loss function. As we aim to maximise the accuracy of the weight ($w(t')$), we use the root-mean-square error (RMSE) as the loss function, calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{t' \in T'} (w(t') - O(t'))^2}{|T'|}} \quad (3)$$

where T' is the set of the opposite context terms from a training dataset, and $O(t')$ is the oracle weight for the term t' within the training dataset. The procedure to obtain $O(t')$ is discussed in Section 4.

4. LEARNING PROCEDURE

This section describes the procedure to derive the weight of each opposite context term to penalise the relevance score of medical records containing the opposite context terms as discussed in Section 3.3. Indeed, Section 4.1 details the set of features $\Phi^{t'}$ that are used to estimate the weight $w(t')$ for the unseen queries. Then, we explain how the estimated effective penalising weight $O(t')$ for an opposite context term t' is obtained in Section 4.2. Finally, Section 4.3 further describes in detail the regression technique (learner) and the objective function (i.e. loss function) that we use to learn the penalising weights.

Table 5: List of features used to predict the penalising weight of the opposite context terms.

Parameter	Description
$Q = t_i...t_n$	query Q of length n contains query terms $t_i...t_n$
T	# terms in the collection
N	# documents in the collection
t	a term t occurring in the query Q
t'	the term t' having a context opposite to the corresponding query term t (i.e. an opposite context term)
$\mathrm{P}(t_1, t_2)$	the maximum likelihood estimation function of the joint probability of any terms t_1 and t_2, estimated as the fraction of documents where they co-occur
$\mathrm{P}(t_1)$	the maximum likelihood estimation function of the term t_1, estimated as the fraction of documents where the term t_1 occurs

Feature types	ID	Definition
term frequency	1	$tf(t)$: raw frequency of term t in the collection
	2	$tf(t')$: raw frequency of term t' in the collection
	3	$log\frac{tf(t)}{N}$: a variant of the term frequency of t
	4	$log\frac{tf(t')}{N}$: a variant of the term frequency of t'
document frequency	5	$df(t)$: # of documents in the collection that contain term t
	6	$df(t')$: # of documents in the collection that contain term t'
	7	$log\frac{T}{df(t)+1}$: a variant of the invert document frequency of t
	8	$log\frac{T}{df(t')+1}$: a variant of the invert document frequency of t'
co-occurrence frequency	9	$\#co\text{-}occur(t', t)$: # of documents containing both terms t and t'
	10	$co\text{-}occur(t', t) = \log \mathrm{P}(t', t)$: a variant of the co-occurrence between terms t and t'
	11	$co\text{-}occur(t', Q) = \sum_{t_i \in Q} \log \mathrm{P}(t', t_i)$: a variant of the co-occurrence between the term t' and other terms in the query Q
	12	$EMIM(t', t) = \log \frac{\mathrm{P}(t', t)}{\mathrm{P}(t') \cdot \mathrm{P}(t)}$: a variant of the co-occurrence between terms t and t'
query length	13	# of the terms in the query (n)

4.1 Learning Features

We firstly identify a set of features $\Phi^{t'}$ of an opposite context term t' to be used to train a learner to identify the penalising weight of the opposite term. These features should correlate with the weight $O(t')$ that could bring about the optimal performance, and are generalised across terms. Table 5 lists the 13 features used in this paper. We focus on features that can be obtained directly from the corpus, which makes our experiments reproducible; however, there may be other features that can be explored in future work.

We focus on 4 types of features, namely term frequency, document frequency, frequency of co-occurrence, and query length. Indeed, the first two types include the classical term and document frequency statistics and their variants (Features 1-8), which model the ubiquity and specificity of a particular term [19]. Specifically, these features consist of the term occurrence statistics of both an opposite context term t' and its corresponding query term t. The higher value of these features, the more discriminative the term is. Indeed, we use the term and the document frequencies of both t and t', since the importance of the opposite context term t' may depend on the discriminative power of both the opposite context term t' and its corresponding query term t. The next set of features (Features 9-12) are related to the co-occurrence frequency. It has been shown that a term that frequently co-occurs with the query terms often relates to the query [5]. Therefore, it is intuitive that the medical records containing the opposite context terms that frequently co-occur with the query terms should not be highly penalised. In particular, Features 9-12 measure the co-occurrence of the opposite context term t' with the query terms using different co-occurrence variants, such as the raw number of documents where the term t' and its corresponding query term t co-occur and the EMIM (Expected Mutual Information Measure) [34] of terms t' and t. Finally, since a long query tends to be more complex and hence more difficult, Feature 13 counts the number of query terms ($|Q|$). Indeed, a long query provides more evidence to infer the relevance of EMRs, and hence it is possible to derive more opposite context terms to penalise the non-relevant medical records.

4.2 Estimating an Effective Penalising Weight

To identify the effective penalising weight of the opposite context terms, we follow [9] and assume the independence between the opposite context terms, when estimating the penalising weight of each opposite context term one at a time, when adding them to a query. Indeed, on the training set, when estimating the effective penalising weight of each opposite context term ($w(t')$ in Equation (1)), we add the opposite context term t' to the query, and identify the oracle weight $O(t')$ of t', which is the weight that provides the highest retrieval effectiveness, in terms of a particular retrieval measure (e.g. MAP or precision at 10), when ranking medical records using a particular ranking model (e.g. BM25). In particular, we sweep the penalising weight between -1 and 1 to find the best penalising weight for each opposite context term t'. We allow the penalising weight to be between -1 and 1, since it is also possible that the occurrences of an opposite context term in a medical record may infer the relevance of a medical record. For example, for a query to find patients with 'hearing loss', the medical record stating "patient presents no signs of hearing" (i.e. represented as "n$hearing") is likely to be relevant. Therefore, in our model, we allow the penalising weight to be either negative or positive, so that the learner will decide based on the term's features what is the effective penalising weight.

4.3 Learning the Penalising Weight

From a training dataset, we have examples of penalising weights for opposite context terms and their corresponding features. We then learn to predict the penalising weight of each unseen opposite context term based on its features. In particular, while any regression technique can be used, we deploy Gradient Boosted Regression Trees (GBRT) [33]

(as implemented in the jforest package [13][1]) as a learner, since it has been shown to be effective and efficient in several search and regression tasks [23, 33]. We use the root-mean-square error (RMSE) as the loss function (i.e. Equation (3)) when learning the penalising weight of an opposite context term. Our proposed framework leverages term frequency, document frequency, and the co-occurrence statistics of the terms in the corpus, introduced in Section 4.1, as learning features for the GBRT learner.

5. EXPERIMENTAL SETUP

We have emphasised the challenge of handling negation language in medical records search and proposed our learning framework to deal with such a challenge in Sections 3 and 4. In this section, we discuss our experimental setup to evaluate the effectiveness of our proposed negation handling framework. In particular, Section 5.1 describes the used medical test collection. Section 5.2 discusses the ranking models used in our experiments. Finally, we discuss the setting of the GBRT learner in Section 5.3.

5.1 Corpus/Queries/Measures

We use the test collection provided by the TREC 2011 [36] and 2012 [35] Medical Records track to evaluate our proposed framework. The task is to identify patient *visits* relevant to a given query topic. Each visit contains all of the medical records associated with a patient's visit to a hospital. Due to privacy concerns [36], a *visit* is used to represent a *patient* as a unit of retrieval. The collection contains medical records from the University of Pittsburgh NLP Repository[2], which consists of approximately 102k medical records. These medical records can be mapped to 17,265 different patient visits. We evaluate our proposed framework using the 34 and 47 topics from the TREC 2011 and 2012 Medical Records track, respectively. Example query topics include:

Q101: Patients with hearing loss
Q102: Patients with complicated GERD who receive endoscopy
Q137: Patients with inflammatory disorders receiving TNF-inhibitor treatments
Q179: Patients taking atypical antipsychotics without a diagnosis schizophrenia or bipolar depression

We evaluate the effectiveness of our proposed framework, using the TREC Medical Records track official measures, namely the bpref [8] and precision at 10 (P10) measures for TREC 2011, and the infAP [38], infNDCG [38] and P10 measures for TREC 2012. bpref, infAP and infNDCG are used since the gold standard judgements are incomplete [35, 36]. The higher these retrieval measures, the more effective the retrieval system. In addition, the paired t-test is used to measure the statistical significance (at $p < 0.05$ and $p < 0.01$) of the difference between the retrieval performances of our proposed framework and each compared baseline.

5.2 Medical Records & Visits Ranking

We index the medical records using the Terrier retrieval platform [27], applying Porter's English stemmer and removing stopwords. The TREC task encompasses ranking patient visits instead of retrieving medical records directly. Indeed, retrieving patient visits having medical records relevant to the query is similar to the expert search task [6], which aims to rank people (e.g. employees within organisations) based on the relevance of documents associated to them. It has been shown in TREC that the existing expert search approaches (e.g. [24, 39]) could effectively handle the TREC medical records search task. Hence, in this work, we also deploy a well-established approach previously developed for expert search –the Voting Model [25]– such that patients are ranked based on their medical records [24, 39]. The Voting Model [25] views patient visits ranking as a voting process. The ranked medical records (denoted $R(Q)$) literally vote for the relevance of their associated patient visits. Specifically, the score of each medical record in $R(Q)$ is used to estimate the relevance of candidate patients by using a voting technique, such as CombMAX and expCombSUM. Indeed, each voting technique firstly ranks medical records based on their relevance towards a query using any traditional document ranking model (e.g. BM25 [28], DFR DPH [2]). Then the relevance scores of medical records are aggregated to define the relevance score of their associated patient visits.

In particular, we use the effective parameter-free DPH term weighting model [2] to rank medical records. Then, to rank the patient visits, we deploy expCombSUM [25], as it focuses more on the highly relevant medical records (i.e. medical records in the top ranks), while voting for the relevance of the patient visits. Specifically, expCombSUM estimates the relevance of a patient visit v with respect to a query Q as follows [24]:

$$score\_visit_{expCombSUM}(v, Q) = \sum_{d \in R(Q) \cap profile(v)} e^{score(d,Q)} \quad (4)$$

where $R(Q) \cap profile(v)$ contains the set of medical records in the ranking $R(Q)$ that are also associated to the patient visit v; $score(d, Q)$ is the relevance score of medical record d for query Q, as obtained by a standard weighting model (namely, DFR DPH). The number of medical records in $R(Q)$ to vote for the relevance of the patient visits is limited to 5,000, as suggested in [24].

5.3 Gradient Boosted Regression Trees

To learn the weight of the opposite context terms (i.e. $w(t')$ in Equation (1)) from the features, we use the default setting of GBRT from the jforest package. Since the number of topics in each topic set is small (i.e. 34 topics for TREC 2011 and 47 topics for TREC 2012), we choose to train the learner using the other topic set, when testing with one particular topic set. For example, we train the GBRT learner on the TREC 2011 topic set, when testing with the TREC 2012 topic set, and vice versa. We refer to this setting as *cross-collection validation* (x-collections). When training the term weight estimation learner, we aim to minimise the root-mean-square error (RMSE) of the weight $w(t')$ of each opposite context term t', as per Equation (3). We train the effective penalising weight based on the achieved retrieval performance, in terms of bpref and infNDCG, when the query topics from the TREC 2011 and TREC 2012 are used as the training topics, respectively.

6. EXPERIMENTAL RESULTS

Next, we discuss the experimental results conducted using our proposed framework in Section 6.1. Moreover, as query expansion techniques have been shown to be effective for the medical records search task [11, 20, 39], we report the results of our framework when a traditional query expansion is also

[1] http://code.google.com/p/jforests/
[2] http://www.dbmi.pitt.edu/nlpfront

Table 6: Retrieval performances of our proposed negation handling framework in comparison to several existing approaches. Statistically significant differences (paired t-test) at $p < 0.05$ and $p < 0.01$ are denoted with symbols, compared to the traditional approach baseline (* and **) and to the post-retrieval filtering baseline ($^{\oplus}$ and $^{\oplus\oplus}$), respectively.

Approach	TREC 2011		TREC 2012		
	bpref	P10	infNDCG	infAP	P10
Traditional approach	0.4871	0.5765	0.4167	0.1703	0.4638
Post-retrieval filtering [29, 37]	0.4477	0.5235	0.3841	0.1577	0.4340
Context Identification [20]	**0.5055**$^{\oplus\oplus}$	**0.5794**$^{\oplus}$	0.4355$^{**,\oplus}$	**0.1833**$^{**,\oplus}$	**0.4894**
LTHN (x-collections)	0.5005$^{\oplus\oplus}$	0.5647	**0.4357**$^{*,\oplus}$	0.1830$^{*,\oplus\oplus}$	0.4830

deployed in Section 6.2. Finally, an ablation study to identify the importance of features is presented in Section 6.3

6.1 The Negation Handling Framework

First, we evaluate the retrieval performance of our proposed framework using the TREC 2011 and 2012 Medical Records track test collections. Specifically, we compare the effectiveness of our proposed negation handling framework with that of the baselines, including the post-retrieval filtering approach [29, 37] (i.e. using DFR DPH to rank medical records and filtering out medical records with opposite context query terms), the context identification [20, 24], and a traditional approach where negation is not explicitly handled (i.e. using DFR DPH to rank medical records)

Table 6 compares the retrieval performance of our proposed framework with the aforementioned baselines, in terms of bpref, infNDCG, infAP, and P10. Firstly, we observe that the *post-retrieval filtering* baseline performs worse than the other approaches reported in this paper. This is likely because this approach simply discards all medical records containing query terms with the opposite context, while it may be possible that some of these medical records are relevant. Next, both the *context identification* approach [20] and our proposed learning framework to handle negation, with the fair cross-collection validation setting (namely, *LTHN (x-collections)*) outperform the *traditional approach* baseline, where the negation is not explicitly handled, for most of the official TREC retrieval measures. The only exception is that *LTHN (x-collections)* does not outperform the *traditional approach* baseline for P10 on TREC 2011. Indeed, on the TREC 2012 topic set, the proposed framework, *LTHN (x-collections)*, significantly outperforms the *traditional approach* (paired t-test, $p < 0.05$), up to 4.6% and 7.4% in terms of infNDCG and infAP, respectively. This shows that to attain an effective retrieval performance, a search system should be able to distinguish between the context of terms. In addition, we find that our proposed framework, *LTHN (x-collections)*, did not improve over the *context identification* baseline (i.e. where only the context identification component of our framework is active). This means that for this setting, the penalising weight estimation component of our framework (introduced in Section 3.3) could not effectively penalise non-relevant medical records. We believe that this is because our framework aims to demote medical records containing query terms with the opposite context from the retrieved ranking list; however, the relevance of the retrieved medical records depends only on the occurrence of a small number of query terms. As the evidence (i.e. query terms) used to retrieve medical records is limited, our proposed framework could not effectively demote potentially non-relevant medical records while retaining the relevant ones at the top ranks. Next, in Section 6.2, we examine if having more evidence (i.e. query terms) to infer the relevance of medical records, our proposed framework could further improve the retrieval performance.

6.2 Applying Query Expansion

As local-statistic [1] and external corpus [12] query expansion (QE) approaches have been shown to be effective for the medical records search task [11, 20, 39], in this section, we apply such approaches to improve the query representation by adding more evidence (i.e. query terms) to the queries. In particular, we expect that if QE expands the query with more evidence (i.e. informative terms) to infer the relevance of medical records, our negation handling framework would effectively demote the non-relevant medical records in the ranking list, and hence improve retrieval performance. Therefore, we improve the representation of the queries by using information from both internal and external corpora. Indeed, we apply the DFR Bo1 model [1] to expand the queries with the top 10 informative terms from the top 3 ranked documents retrieved from the medical records collection of the TREC Medical Records and the MEDLINE abstract collection of the TREC 2005 Genomics [15] tracks.

Table 7 compares the retrieval performance, after applying the aforementioned QE technique on both our proposed framework and the baselines. In particular, we use the same baselines (i.e. the traditional, the post-retrieval filtering, and the context identification approaches) and the same retrieval effectiveness measures (i.e. bpref and P10 for TREC 2011, and infNDCG, infAP, and P10 for TREC 2012) as in Section 6.1. In addition, the highest retrieval performance that our framework could achieve is also discussed (i.e. when using the oracle $w(t') = O(t')$).

From Table 7, we firstly observe that after applying QE, the retrieval performances of our proposed framework (namely, *LTHN (x-collections)+QE*) and all of the baselines increase markedly. This shows that, overall, the QE technique could expand the queries with informative terms. Next, as expected, we find that after applying QE, our proposed framework with the cross-collection validation setting, *LTHN (x-collections)+QE*, further improves the retrieval performance. Indeed, our negation handling framework outperforms all of the baselines for all of the reported measures. For TREC 2011, the proposed framework, *LTHN (x-collections)+QE*, performs significantly (paired t-test, $p < 0.05$) better than the traditional approach where the negation is not explicitly handled (*traditional approach+QE*), in terms of bpref (0.5786 versus 0.5567), while the performance, in terms of precision at 10, improves from 0.6527 to 0.6647. For the TREC 2012 topic set, *LTHN (x-collections)+QE* significantly ($p < 0.05$) outperforms the *traditional approach+QE* baseline, for all the reported retrieval measures. In particular, the proposed negation handling framework, *LTHN (x-collections)+QE*, outperforms the traditional baseline (*tra-*

Table 7: Retrieval performances of our proposed negation handling framework in comparison to several existing approaches, after applying query expansion. Statistically significant differences (paired t-test) at $p < 0.05$ and $p < 0.01$ are denoted with symbols, compared to the traditional approach baseline (* and **), to the post-retrieval filtering baseline ($^{\oplus}$ and $^{\oplus\oplus}$), to the context identification baseline ($^{\circ}$ and $^{\circ\circ}$), and to our LTHN framework with the best possible oracle setting ($^{\bullet}$ and $^{\bullet\bullet}$), respectively.

Approach	2011		2012		
	bpref	P10	infNDCG	infAP	P10
Traditional approach+QE	0.5569	0.6529	0.4619	0.1982	0.4702
Post-retrieval filtering+QE	0.5096	0.5735	0.4238	0.1858	0.4681
Context Identification+QE	$0.5733^{\oplus}$	$0.6559^{\oplus}$	$0.4838^{*,\oplus}$	0.2127^{*}	0.5149^{*}
LTHN (x-collections)+QE	$\mathbf{0.5786}^{*,\oplus\oplus}$	$\mathbf{0.6647}^{\oplus\oplus}$	$\mathbf{0.4911}^{*,\oplus,\circ}$	$\mathbf{0.2157}^{*,\oplus,\bullet}$	$\mathbf{0.5362}^{*}$
LTHN (oracle)+QE	$0.5891^{**,\oplus\oplus,\circ}$	$0.6647^{\oplus\oplus}$	$0.5004^{**,\oplus\oplus,\circ}$	$0.2229^{**,\oplus,\circ\circ}$	$0.5447^{**,\oplus}$

ditional approach+QE) by 6.3%, 8.8%, and 14%, in terms of infNDCG, infAP, and P10, respectively. This confirms that the third component of our LTHN framework, namely the penalising weight estimation component, can effectively penalise non-relevant medical records, when several informative terms are used in a query. In addition, we find that with the cross-collection validation setting, our proposed *LTHN (x-collections)+QE* could perform comparably to the best possible setting (i.e. when $w(t') = O(t')$), *LTHN (oracle)+QE*. This shows that our learning framework is robust and could be generalised between the training and test topic sets.

6.3 Feature Importance: Ablation Study

Next, in order to examine the importance of each proposed feature, we conduct an ablation study. Indeed, using the same experimental setup as discussed in Section 6.2, we remove each of the features from the feature space to examine the impact of each feature on the retrieval performance. For example, when we evaluate the importance of Feature 1, we remove Feature 1 from the feature space, while keeping the other 12 features. The increase or reduction in the achieved retrieval effectiveness is an indicator of the importance of the feature. Specifically, the retrieval performance decreases when removing an effective feature, while the performance remains the same or increases when removing a non-effective feature. We compare the importance of each feature based on its impact on the retrieval performance, in terms of bpref and infNDCG for TREC 2011 and 2012, respectively. These two measures are selected as representatives because they are the measures that we use to train the model to estimate the effective penalising term weights (i.e. the weight that could bring about the highest retrieval performance for a particular term), as discussed in Section 5.3. The percentage improvement or reduction, in terms of each retrieval measure, between when a particular feature is removed and when all the features are considered, are summed up to measure the overall impact of that feature. Then, we normalise this measure for each feature by dividing it with that of the most important feature (i.e. the feature with the most negative impact on retrieval effectiveness), in order to easily rank the importance of different features. We refer to the normalised measure as *feature importance*. Hence, the feature whose removal *degrades effectiveness most* has the *highest feature importance*. If the feature importance of a particular feature is zero, the feature has no impact on the retrieval performance, while a negative feature importance indicates that the feature is not useful and can be removed from the feature set.

Figure 1 compares the feature importance of each feature in our feature space. We observe that Features 11, 8 and 7,

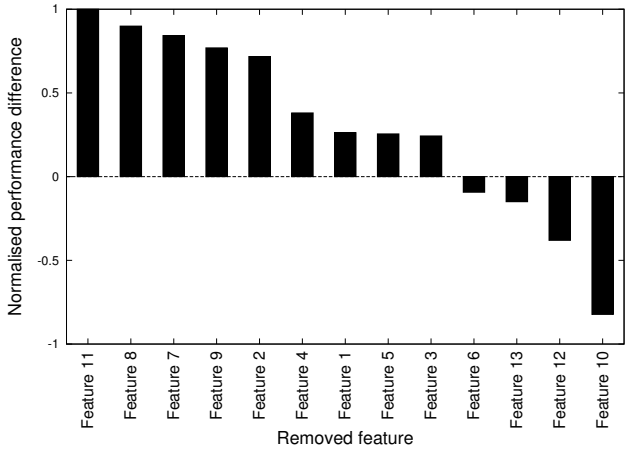

Figure 1: Ablation study of feature importance across both TREC 2011 and 2012

.

which are the co-occurrence between the opposite context term t' and the query Q, an IDF variant of the opposite context term t', and an IDF variant of the query term t corresponding to the opposite context term t', respectively, are the most important features. This is intuitive as the co-occurrence of the opposite context term t' and the terms in the query Q could be used to measure the relatedness between the opposite context term and the query, while the IDF variant of the opposite context term t' and its corresponding query term t could measure the informativeness of both associated terms. In contrast, Features 10, 12 and 13, namely the two variants of the co-occurrence frequency between the opposite context term t' and its corresponding query term t, and the query length, respectively, are the least importance features. Indeed, adding these features to the feature set has a negative impact on retrieval effectiveness. Overall, we find that 9 out of our proposed 13 features are beneficial to obtaining an effective estimation of the penalising weights for the opposite context terms.

7. REPRESENTATION COMBINATION

It has previously been shown that an effective retrieval performance of a medical records search system can be obtained by combing the relevance scores of a term-based and conceptual-based representations [21, 23, 31]. Hence, we examine whether our proposed framework can further improve retrieval performance when combined with a conceptual-based representation, hence resulting in an improved retrieval performance. The conceptual representation approach represents medical records and queries in terms of concepts,

Table 8: Retrieval performances when combining the representation of our proposed framework and a conceptual representation approach. Statistically significant differences (paired t-test) at $p < 0.05$ and $p < 0.01$ are denoted with symbols, compared to when using only the conceptual representation approach (* and **), and when using only our proposed framework with the cross-collection setting ($^{\oplus}$ and $^{\oplus\oplus}$), respectively.

TREC 2011			TREC 2012			
Approach	bpref	P10	Approach	infNDCG	infAP	P10
Conceptual representation [21]	0.5282	0.5294	Conceptual representation [21]	0.4530	0.2126	0.4936
LTHN (x-collections)+QE	0.5786	**0.6647****	LTHN (x-collections)+QE	0.4911	0.2157	0.5362
Representation combination	**0.5809***	0.6294**	Representation combination	**0.5299**$^{**,\oplus}$	**0.2460***	**0.5723**$^{*,\oplus}$
CengageM11R3	0.5520	0.6560	udelSUM	0.5780	0.2860	0.5920
SCIAMED7	0.5520	0.6030	sennamed2	0.5470	0.2750	0.5570
UTDHLTCIR	0.5450	0.6030	atigeo1	0.5240	0.2240	0.5190

instead of terms [21, 22, 31]. For example, *'cerebrovascular accident'*, *'stroke'*, and *'CVA'* are represented with the same concept, as they share a particular conceptual meaning. This could help to reduce the mismatch between terms in a medical records and a query. In this work, we follow Limsopatham et al. [21, 23] and deploy MetaMap [3] – a medical concept recognition tool – to identify concepts in medical records and queries, and represent them in the form of the UMLS Concept Unique Identifier (CUI)[3]. In addition, we also apply the Bo1 QE model to expand the conceptual query with the top 10 informative concepts from the top 3 ranked medical records, retrieved from the TREC Medical Records track's collection.

To combine the representations, we linearly combine the relevance scores of a medical record d towards a query Q, calculated using both our framework ($score_{LTHN}$) and the conceptual representation approach ($score_{conceptual}$), as follows [31]:

$$score(d,Q) = \delta \cdot score_{LTHN}(d,Q) + score_{conceptual}(d,Q) \quad (5)$$

where δ is a parameter to emphasise the relevance score computed using our LTHN framework, which represents medical records and queries based on terms. We set δ to 2.00, as suggested in [21, 31].

Table 8 reports the retrieval performances on the TREC 2011 and 2012 Medical Records track of the representation combination approach (i.e. *representation combination*), which combines the relevance scores of the aforementioned conceptual representation approach, and our LTHN framework after applying the query expansion strategy, discussed in Section 6.2 (namely, *LTHN (x-collections)+QE*). In particular, we compare the retrieval performance of the *representation combination* with the effectiveness of using only either the conceptual representation or our negation handling framework. Moreover, the retrieval performances of the best TREC 2011 and TREC 2012 Medical Records track systems are also reported.

From Table 8, we firstly observe that our proposed framework (*LTHN (x-collections)+QE*) outperforms the conceptual representation alone (namely, *conceptual representation*) for all reported retrieval measures. For example, in terms of precision at 10, our framework significantly (paired t-test, $p < 0.01$) outperforms the *conceptual representation* approach by up to 25.6%. Next, we find that the representation combination approach, which combines the relevance scores computed using our LTHN framework and the conceptual representation approach, could further boost the retrieval performance. Indeed, the achieved retrieval performances are significantly better than both the conceptual representation and our proposed framework (*LTHN (x-collections)+QE*) for almost all the retrieval measures, except for the precision at 10 on the TREC 2011 topic set. Specifically, in terms of bpref, the *representation combination* outperforms the conceptual representation baseline by 10% (paired t-test, $p < 0.05$), and *LTHN (x-collections)+QE* by 0.4%. For TREC 2012, the representation combination achieves an infNDCG of 0.5299, which is significantly better than the conceptual representation baseline ($p < 0.01$) and *LTHN (x-collections)+QE* ($p < 0.05$) by up to 17% and 7.9%, respectively. In terms of infAP, the representation combination approach performs 15.7% significantly ($p < 0.05$) better than the *conceptual representation*, and 14% higher than *LTHN (x-collections)+QE*. In addition, the precision at 10 on TREC 2012 of the *representation combination* is significantly better than both the *conceptual representation* and *LTHN (x-collections)+QE* by up to 15.9% and 6.7%, respectively. These improved results support the conclusion that using the representation combination approach [31], our supervised learning framework to handle negation combines effectively with the conceptual representation approach. Indeed, the attained retrieval effectiveness is comparable to those of the best TREC 2011 and TREC 2012 systems, without resorting to other common approaches for medical records search (e.g. leveraging the semantic relationships of medical terms to infer the relevance of medical records).

[3]We leave as future work the integration of our negation handling approach within conceptual representation.

8. CONCLUSIONS

We have motivated the need for a medical records search system to handle negated language, which is commonly used in the medical domain. We introduced our proposed framework to handle negation by distinguishing between the contexts of terms before their processing, and demoting medical records containing the query terms with the opposite context of the query's intent. Specifically, our framework prevents non-relevant medical records from being ranked highly, by demoting those containing occurrences of opposite context terms (i.e. a term having the opposite context to its corresponding query term). We deploy a supervised learning approach to effectively estimate the penalising weight for the opposite context query terms. In particular, we use the Gradient Boosted Regression Trees (GBRT) to learn the penalising weight of an opposite context term, using features such as term and document frequencies.

We evaluate our proposed framework using the standard test collection from the TREC 2011 and 2012 Medical Records track. Our experimental results show that the proposed framework significantly outperforms several strong baselines. Moreover, our proposed learning framework is effective, as

the retrieval performance achieved using a fair setting (i.e. cross-collection validation) is comparable to that of the best possible setting (i.e. the oracle setting, when $w(t') = O(t')$). In addition, our proposed negation handling framework works effectively with an existing QE approach and could effectively combine with a conceptual representation, using an existing representation combination approach. Specifically, the achieved retrieval performance is comparable to the state-of-the-art results among participants in the TREC 2011 and TREC 2012 Medical Records track, while these top-performing systems also deploy approaches to handle other challenges in the medical records search, such as using the semantic relationship of terms to reformulate the query.

For future work, we aim to integrate our proposed negation handling framework within a reasoning model that infers the relevance of a medical record based on the presence or absence of medical conditions associated with the query.

9. REFERENCES

[1] G. Amati. Probabilistic Models for Information Retrieval based on Divergence from Randomness. *PhD thesis*. University of Glasgow, 2003.

[2] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog Track. In *Proc. of TREC*, 2007.

[3] A. R. Aronson and F. Lang. An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.*, 17(3), 2010.

[4] M. Averbuch, T. Karson, B. Ben-Ami, O. Maimond, and L. Rokachd. Context-sensitive medical information retrieval. In *Proc. of AMACL*, 2003.

[5] J. Bai, J. Y. Nie, G. Cao, H. Bouchard. Using query contexts in information retrieval. In *Proc. of SIGIR*, 2007.

[6] K. Balog, P. Thomas, N. Craswell, I. Soboroff, and P. Bailey. Overview of the TREC 2008 Enterprise Track. In *Proc. of TREC*, 2008.

[7] M. Bendersky, D. Metzler, and W. B. Croft. Parameterized Concept Weighting in Verbose Queries. In *Proc. of SIGIR*, 2011.

[8] C. Buckley and E. Voorhees. Retrieval Evaluation with Incomplete Information. In *Proc. of SIGIR*, 2004.

[9] G. Cao, J.Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proc. of SIGIR*, 2008.

[10] W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. Biomed. Inform.*, 34(5), 2001.

[11] D. Demner-Fushman, S. Abhyankar, A. Jimeno-Yepes, R. Loane, B. Rance, F. Lang, N. Ide, E. Apostolova, and A.R. Aronson, A Knowledge-Based Approach to Medical Records Retrieval. In *Proc. of TREC*, 2011.

[12] F. Diaz and D. Metzler. Improving the estimation of relevance models using large external corpora. In *Proc. of SIGIR*, 2006.

[13] Y. Ganjisaffar, R. Caruana, and C. V. Lopes. Bagging gradient-boosted trees for high precision, low variance ranking models In *Proc. of SIGIR*, 2011.

[14] W. Hersh. *Information retrieval: A health and biomedical perspective (3rd ed.)*. New York : Springer, 2009.

[15] W. Hersh, A. Cohen, J. Yang, R. Bhupatiraju, and M. Hearst. TREC 2005 Genomics Track Overview. In *Proc. of TREC*, 2005.

[16] H. Jain, C. Thao, and Z. Huimin. Enhancing electronic medical record retrieval through semantic query expansion. *Inf. Syst. E-Business Management*, 10(2), 2012.

[17] B. King, L. Wang, I. Provalov, and J. Zhou. Cengage Learning at TREC 2011 Medical Track. In *Proc. of TREC*, 2011.

[18] B. Koopman, P. Bruza, L. Sitbon and M. Lawley. Analysis of the effect of negation on information retrieval of medical data. In *Proc. of ADCS*, 2010.

[19] M. Lease, J. Allan, and B. Croft. Regression Rank: Learning to Meet the Opportunity of Descriptive Queries. In *Proc. of ECIR*, 2009.

[20] N. Limsopatham, C. Macdonald, R. McCreadie, and I. Ounis. Exploiting Term Dependence while Handling Negation in Medical Search. In *Proc. of SIGIR*, 2012.

[21] N. Limsopatham, C. Macdonald, and I. Ounis. A Task-Specific Query and Document Representation for Medical Records Search. In *Proc. of ECIR*, 2013.

[22] N. Limsopatham, C. Macdonald, and I. Ounis. Inferring Conceptual Relationships to Improve Medical Records Search. In *OAIR* 2013.

[23] N. Limsopatham, C. Macdonald, and I. Ounis. Learning to Combine Representations for Medical Records Search. In *SIGIR* 2013.

[24] N. Limsopatham, C. Macdonald, I. Ounis, G. McDonald, and M. Bouamrane. University of Glasgow at Medical Records track 2011: Experiments with Terrier. In *Proc. of TREC*, 2011.

[25] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *Proc. of CIKM*, 2006.

[26] P. Mutalik, A. Deshpande, and P. Nadkarni. Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study using the UMLS. *J. Am. Med. Inform. Assoc.*, 6, 2001.

[27] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proc. of OSIR at SIGIR*, 2006.

[28] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proc. of SIGIR*, 1994.

[29] G. Salton and M. McGill. Introduction to Modern Information Retrieval. *McGraw-Hill, Inc.*, New York, NY, USA, 1986.

[30] E. L. Siegel and D. S. Channin. Integrating the healthcare enterprise: A primer. Part 1. Introduction. *RadioGraphics 2001*, 21(5), 2001.

[31] P. Srinivasan. Optimal document-indexing vocabulary for MEDLINE. *Inf. Process. Manage.*, 32(5), 1996.

[32] E. Tambouris, M.H. Willimas MH, and C. Makropoulos. Co-operative health information networks in Europe: experience from Greece and Scotland. *J. Med. Internet. Res.*, 2(2):e11, 2000.

[33] S. Tyree, K. Q. Weinberger, K. Agrawal, and J. Paykin. Parallel boosted regression trees for web search ranking. In *Proc. of WWW*, 2011.

[34] C. J. van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. *J. Doc.*, 33(2), 1977.

[35] E. Voorhees and W. Hersh. Overview of the TREC 2012 Medical Records Track. In *Proc. of TREC*, 2012.

[36] E. Voorhees and R. Tong. Overview of the TREC 2011 Medical Records Track. In *Proc. of TREC*, 2011.

[37] D. Widdows. Orthogonal negation in vector spaces for modelling word-meanings and document retrieval. In *Proc. of ACL*, 2003.

[38] E. Yilmaz, E. Kanoulas, and J. A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *Proc. of SIGIR*, 2006.

[39] D. Zhu and B. Carterette. Combining Multi-level Evidence for Medical Record Retrieval. In *Proc. of SHB*, 2012.