Automatic Generation of a Qualified Medical Knowledge Graph and Its Usage for Retrieving Patient Cohorts from Electronic Medical Records

Conference Paper · September 2013

DOI: 10.1109/ICSC.2013.68

Travis Goodwin
National Institutes of Health

2013 IEEE Seventh International Conference on Semantic Computing

Automatic Generation of a Qualified Medical Knowledge Graph and its Usage for Retrieving Patient Cohorts from Electronic Medical Records

Travis Goodwin
Human Language Technology Research Institute
University of Texas at Dallas
Dallas, Texas 75080
Email: travis@hlt.utdallas.edu

Sanda M. Harabagiu
Human Language Technology Research Institute
University of Texas at Dallas
Dallas, Texas 75080
Email: sanda@hlt.utdallas.edu

***Abstract*—An extraordinary amount of clinical information is available within Electronic Medical Records. However, interpreting this knowledge typically demands a significant level of clinical understanding. This can facilitated by access to structured knowledge bases. However, even if vast, biomedical knowledge bases have very limited relational information available. In contrast, clinical text expresses many relations between concepts using an extraordinary amount of variation regarding the author's belief state – whether a medical concept is present, uncertain, or absent. In this paper, we propose a method for automatically constructing a graph of clinically related concepts based on their belief state. For this purpose, we first devise a method for classifying the belief state of certain medical concepts. Second, we designed a technique for constructing a graph of related medical concepts qualified by the physician's belief value. Thirdly, we demonstrate several techniques for inferring the similarity between qualified medical concepts, and present a generalized algorithm for determining the second-order similarity between qualified medical concepts. Finally, we show that incorporating the knowledge encoded from this graph yield competitive results when applied to query expansion for the retrieval of hospital patient cohorts.**

I. Introduction

A. The Problem

More and more clinical data is available through massive warehouses of Electronic Medical Records (EMRs). Hospitals throughout the United States and other countries process millions of EMRs annually. The notes within EMRs typically include a variety of clinical information, including medical history, physical exam findings, lab reports, radiology reports, operative reports, as well as discharge summaries. Information about a patient's medical problems, treatments, and clinical course is also available from EMRs. This information is essential for conducting comparative effectiveness research, defined in a brief report from the National Institute of Medicine published in 2009 as the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve delivery of care [1].

No.	Topic
156	Patients with depression on anti-depressant medication.
160	Patients with low back pain who had imaging studies.
172	Patients with peripheral neuropathy and edema.
184	Patients with colon cancer who had chemotherapy.

TABLE I: Examples of topics provided as part of the TRECMed evaluation. The numbers correspond to the topic numbers evaluated in TRECMed.

Because EMRs do not document the rationale for medical decisions, patient cohort studies need to be undertaken for understanding the progression of disease as well as the factors that influence clinical outcomes. Patient cohort identification has been the target of an information retrieval (IR) challenge task performed in the Text REtrieval Conference (TREC) in 2011 and 2012, under the medical records track (TRECMed) [2], [3]. The TRECMed organizers aimed to develop a retrieval problem pertinent to real-world clinical medicine by (a) enabling access to a large corpus of de-identified EMRs available from the University of Pittsburgh Medical Center and (b) a set of 85 retrieval queries called topics, reflecting patient filtering criteria similar to those specified for participation in clinical studies and the list of findings conditions developed by the National Institute of Medicine.

This retrieval task considered 95,703 de-identified EMRs which were generated from multiple hospitals during 2007. The EMRs were grouped into hospital visits consisting of one or more medical reports from each patient's hospital stay. Thus, the EMRs were organized into 17,199 different patient hospital visits. Each visit had the patient's admission diagnoses, discharge diagnoses, and related ICD-9 codes. When retrieving a patient cohort relevant to a query, a system produced a ranked list of hospital visits, according to the relevance model that was considered.

The 35 topics evaluated in 2011 and the 50 topics evaluated in 2012 were characterized by (a) usage of medical concepts (e.g. acute coronary syndrome or plavix) and (b) constraints imposed on the patient population (e.g. children, female patients). A subset of the topics is illustrated in Table I.

To be able to satisfy the information need expressed within

978-0-7695-5119-7/19 $26.00 © 5119 IEEE
DOI 10.1109/ICSC.2013.68

IEEE computer society

the topics, the capability of recognizing medical knowledge in the form of medical conditions, symptoms, treatments, or tests is needed. Furthermore, the ability to satisfy the constraints expressed in the topics was also essential. However, an important barrier to accurately retrieving patient cohorts is due to the way in which physicians write about medical concepts: they often use hedging or linguistic means of expressing an opinion, rather than a fact. Medical science involves asking hypotheses, experimenting with treatments, and reasoning from medical evidence. Consequently, clinical writing reflects this modus operandus with a rich set of speculative statements.

By taking this observation into account, we decided to explore a knowledge representation that (1) takes into account the physician's degrees of belief – qualifications of the medical concepts mentioned in EMRs; and that (2) can be acquired automatically from a large corpus of EMRs. Our work considers that all medical concepts within an EMR fall within the categories of (1) medical problems (e.g. LUNG CANCER), (2) medical tests (e.g. CT – indicating an X-ray computed tomography scan, or (3) medical treatments (e.g. TYLENOL). In order to capture the belief values that physicians express with regards to medical concepts, we have considered (a) six types of assertions[1] that were used to qualify the state of a patient's medical problem in the 2010 i2b2/VA challenge; (b) three additional assertions that qualify a patient's treatments, (c) an assertion that applies onto medical tests, and (d) a new assertion that applies to medical problems, treatments, and tests. This classification follows the framework devised in the 2010 i2b2/VA challenge [4], which tasked participants with categorizing medical concepts as problems, treatments, or tests and with classifying the assertion for each medical concept. Table II lists the assertions that we considered for qualifying the physician's belief status.

Clearly, belief values associated with medical concepts mined from EMRs encode a new form of semantic knowledge which can enable several forms of reasoning. In this paper, we focus on organizing this knowledge such that it can be be useful for retrieving relevant patient cohorts, given a topic similar to those used in the TRECMed evaluation. By capturing the assertions associated with medical concepts, we are able to build a novel form of medical knowledge, which we call "qualified medical knowledge." We organize this knowledge into a graph, which we call the Qualified Medical Knowledge Graph (QMKG) which consists of (a) vertices which are triples of the form [lexical medical concept, medical concept type, assertion] (e.g. ["atrial fibrillation," PROBLEM, PRESENT]); and (b) weighted edges which indicate cohesive strengths between their associated vertices, as indicated in EMRs. Moreover, in order to ascertain the quality of the QMKG, we performed both an intrinsic and an extrinsic evaluation. We evaluate the intrinsic quality of the QMKG by comparing it to the graph underlying the Unified Medical Language System, which encodes over 2 million biomedical concepts, which are each assigned a concept unique identifier (CUI), and 54 relationships types between encoded concept. We evaluate the extrinsic utility of the graph by using the QMKG to improve the quality of patient cohort retrieval by enabling a method for query expansion that is based on the weighted structure of the QMKG. For this purpose, we make use of the patient cohort retrieval system reported in [5], [6], which was developed for evaluations in TRECMed 2011 and 2012.

Traditionally, retrieval models do not take into account belief values asserted about concepts. Moreover, semantic information such as word senses has been shown to not improve the accuracy or completeness of retrieval results. This is because semantic information that is too fine-grained seems not to be beneficial to retrieval quality. For example, Voorhees used WordNet [7] as a tool for query expansion [8] with the TREC collection [9]. By expanding query terms with WordNet synonyms, hypernyms, or hyponemys, documents which were retrieved using the SMART retrieval system [10] were more relevant only when queries were short. But, when WordNet was used for word sense disambiguation of the documents, [11], retrieval performance was in fact degraded.

Taking into account these lessons learned, we investigated if, for the problem of patient cohort retrieval, the recognition of medically significant semantic information in the query topics could improve retrieval by informing query expansion methods.

B. The Approach

In this paper, we present an automatic method for generating the QMKG by processing a large corpus of EMRs. The nodes of this graph are triplets of (1) lexicalized medical concepts (e.g. PNEUMONIA), (2) the associated medical concept type (e.g. PROBLEM), and (3) the belief value held by the author of the EMR concerning the associated medical concept (e.g. HYPOTHETICAL). The content of each node of the graph is provided by (i) an automatic method of identifying medical concepts in EMRs, and (ii) an automatic method of asserting the belief value of the respective medical concept. Figure 1 illustrates sentences from EMRs with their associated vertices in the QMKG.

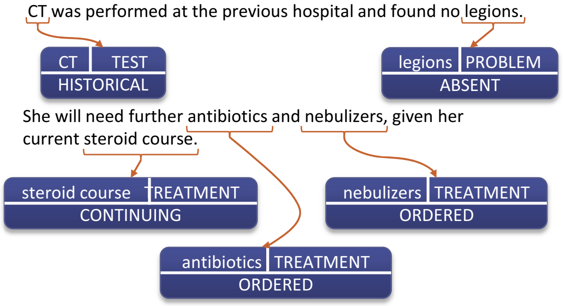
CT was performed at the previous hospital and found no legions.
CT
TEST
HISTORICAL
legions
PROBLEM
ABSENT
She will need further antibiotics and nebulizers, given her current steroid course.
steroid course
TREATMENT
CONTINUING
nebulizers
TREATMENT
ORDERED
antibiotics
TREATMENT
ORDERED

Fig. 1: Examples of medical concepts and their associated graph nodes.

An edge between two graph nodes exists if the corresponding medical concepts co-occur within a window of λ tokens (for

[1] In this paper, we refer to assertions and belief values interchangeably. An "assertion" is considered to be one possible belief value.

Assertion Value	Problem	Treatment	Test	Definition	EMR Excerpt
HISTORICAL	✓	✓	✓	The indicated medical concept occurred during a previous hospital visit.	…the patient's past medical history is significant for CONGESTIVE HEART FAILURE…
CONDITIONAL	✓	✓	✓	The mention of the indicated medical concept asserts that it occurs only during certain conditions.	…[we will] likely readmit him for REHAB once the WOUND has HEALED and [a] proper PROSTHETIC FITTING can be achieved.…
PRESCRIBED		✓		The indicated treatment has been assigned and will begin sometime after this moment.	…she was given ROCEPHIN and ZITHROMAX…
ABSENT	✓	✓	✓	The note asserts that the indicated medical concept does not exist at this moment.	…the patient denies any CHEST PAIN at this time…
SUGGESTED		✓		The indicated treatment or test is advised, though it cannot be assumed to actually occur.	…it was recommended that he be on ALLOPURINOL long-term…
PRESENT	✓			The indicated problem is still active at this moment.	…there is a moderate PERICARDIAL EFFUSION…
HYPOTHETICAL	✓			The note asserts the patient may develop the indicated problem.	…she is to return for any WORSENING PAIN, FEVERS, or PERSISTENT VOMITING…
ORDERED			✓	The indicated treatment has been scheduled and will be completed sometime after this moment.	…we will do a PULMONARY FUNCTION TEST with DESATURATION STUDY…
ASSOCIATED WITH ANOTHER	✓			The mention of the medical problem is associated with someone other than the patient.	…father died of LUNG CANCER probably related to ASBESTOS EXPOSURE…
POSSIBLE	✓			The note asserts that the patient may have a problem, but there is some degree of uncertainty.	…SHORTNESS OF BREATH: I believe that this may represent worsening for PULMONARY HYPTERTENSION, but it could also be secondary to an increase in her EFFUSIONS…
ONGOING	✓	✓		The indicated problem or treatment persists beyond this moment.	…as per nephrology, continue DIALYSIS…
CONDUCTED			✓	The indicated medical test been performed and completed as of this moment.	…a PERIPHERAL IV was placed in the right AC and UNASYN 3 GRAMS IV was given…

TABLE II: Assertion values for the medical concepts (typeset in SMALLCAPS) extracted in each excerpt (excerpts were selected such that all medical concepts share the same assertion value). In this table, a "moment" refers to the specific instant in time in which the particular medical concept was written.

our experiments, we set $\lambda = 20$) within the same EMR. This idea was inspired by the SympGraph methodology reported in [12] which models symptom relationships in clinical notes. The medical concepts that we recognize comprise, in addition to symptoms, diseases, injuries, and other types of concepts that represent a medical problem. In addition, the graph also encodes treatments and medical tests within vertices. The co-occurrence relations indicate links between the graph-nodes and these edges learn non-uniform weights that dictate the way in which a query topic is expanded. Figure 2 illustrates a portion of the QMKG and the way in which it can be used to inform the query expansion for a patient cohort topic.

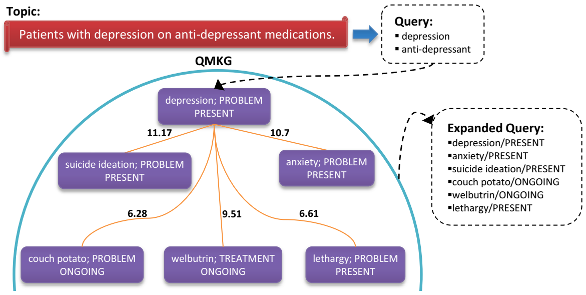
Topic:
Patients with depression on anti-depressant medications.
Query:
▪ depression
▪ anti-depressant
QMKG
depression; PROBLEM PRESENT
11.17
10.7
suicide ideation; PROBLEM PRESENT
anxiety; PROBLEM PRESENT
6.28
9.51
6.61
couch potato; PROBLEM ONGOING
welbutrin; TREATMENT ONGOING
lethargy; PROBLEM PRESENT
Expanded Query:
▪depression/PRESENT
▪anxiety/PRESENT
▪suicide ideation/PRESENT
▪couch potato/ONGOING
▪welbutrin/ONGOING
▪lethargy/PRESENT

Fig. 2: Example of query expansion terms provided by our concept graph for a given topic.

The query terms DEPRESSION and ATYPICAL ANTIDEPRESSANT are each mapped to their associated vertices in the QMKG. For simplicity, we only illustrate the keyword DEPRESSION, which is mapped to the node (*depression*, PROBLEM, PRESENT). This node is connected to a set of related concepts indicated by its neighboring nodes. To be able to expand queries similar to the one illustrated in Figure 2, we focused on learning the weights of the edges within the QMKG. We did this because the selection of expanded terms for enhancing a query is based on the assumption that the weight connecting a vertex to its neighbors indicates the strength of the relationship between those concepts, and thus relative utility of a potential expansion. Moreover, due to the fact that the same medical concept, when qualified by different belief states (assertions), will correspond to different vertices which will in turn have varying weights on their edges. This allows us to discover new correlations between medical concepts encoded in UMLS which are not currently connected in UMLS, and furthermore, to order the strength of these connections under different belief qualifications.

In addition, we have evaluated the impact of the QMKG in expanding queries, and thus generating much improved patient cohort retrieval results. We use our cohort identification system [5], [6] used for the evaluation of the TRECMed challenges in 2011 and 2012. We find that the queries expanded based on the graph we present in this paper produce cohorts that are 23.7% more accurate than those obtained without access to the information encoded within the graph. To learn the weights of the co-occurring links, we have considered 4 different techniques and found that the best results were obtained when the PMI similarity measure was used. Furthermore, we believe that our medical concept/belief graph can also be used for learning how to best rank the patient cohorts and to rely on the feedback from medical experts. Like in [12], the graph can dynamically updated when new EMRs are considered.

The remainder of this paper is structured as follows. Section 2 illustrates the process of automatically identifying medical concepts, their concept type, and their belief values from EMRs. Each medical concept, along with associated concept type and assertion value becomes a node in our qualified medical knowledge graph (QMKG). Section 3 details the way in which weights of the QMKG edges are learned while Section 4 presents the evaluation of our assertion classification and the utility of QMKG when applied to patient cohort retrieval. Section 5 summarizes the conclusions.

II. Identifying Medical Concepts and Their Assertions

The ability to retrieve patient cohorts relevant to topics tested in the TRECMed challenges requires (a) the capability to interpret the topic and (b) the capability to use that interpretation in the processing of a patient's hospital visits. As all the topics used some form of medical knowledge, the first research question focused on the forms of medical knowledge that is most adequate for such interpretation. As the topics were expressed in natural language, the first choice was to consider a resource where lexico-semantic medical knowledge is encoded, such as the Unified Medical Language System (UMLS)[2] [13]. Open-source software, such as MetaMap [14] or, more recently, cTakes [15] can parse the topics and the EMRs to assign concept unique identifiers (CUI) which correspond to entries in UMLS. However, the semantic network available from UMLS involves a large set of concepts that were organized by ontological principles, rather than the latent semantics that can be derived from the large corpus of EMRs. In order to decide on the conceptual representation, we also considered the more general framework developed by the i2b2 challenge in 2010 [4]. The object of this framework is to identify medical concepts in clinical texts and moreover to assign several possible values to capture the degree of belief associated with the medical concepts. Because so many lexico-semantic resources exist for processing clinical texts, i2b2 proposed a challenge to find which resources and which features produce the best results for recognizing medical concepts. But, more importantly, the 2010 i2b2 challenge brought to the forefront of research in medical informatics the problem that recognizing medical concepts alone is not sufficient. When medical concepts are used in clinical documents, physicians also express assertions about those concepts, namely that a medical problem is present or absent, that a treatment is conditional on a test, or that the clinician is uncertain about a medical concept. The i2b2 2010 challenge considered assertions only for medical problems. In our research, for retrieving patient cohorts, we have extended the problem of assertion classification in two ways: first we have produced assertion (or belief values) for all medical concepts that we have automatically identified; second, we have considered 6 additional values which are defined in Table II.

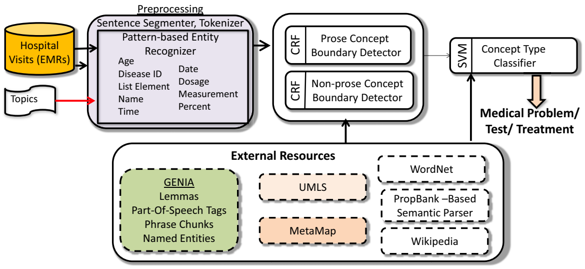

Fig. 3: Concepts identification

A. Medical Concept Recognition

Medical concepts in the form of (1) medical problems, such as ATRIAL FIBRILLATION (irregular heart beat); (2) treatments, such as ABLATION (removal of undesired tissue); and (3) tests, such as ECG (electrocardiogram) were recognized using the methods reported in [16]. This method recognizes medical concepts in two steps:

Step 1: Identification of the boundaries within text that refers to a medical concept;

Step 2: Classification of the medical concept into (a) medical problems, (b) medical treatments, or (c) medical tests.

In Step 1, medical concepts in EMRs were detected both within the narrative of the report and within the structured fields (e.g. chief complaint), and thus two different classifiers implemented as Conditional Random Fields (CRF) were trained on the i2b2 annotations and using on the TRECMed documents to extract medical concepts. As illustrated in Figure 3, recognition of medical concepts benefits from several lexico-semantic resources. In addition to UMLS [17] and MetaMap [14], the Genia annotations [18] used for Biomedical text mining were incorporated.

Lexico-semantic information, especially aimed at identifying lemmas and multi-word expressions, was mined from WordNet. Additional information about lemmas, part-of-speech, and phrasal chunks as well as names of entities was provided by the Genia tools. Concept type information from Wikipedia was also used. All these features informed not only the concept boundary classifiers, but also the concept-type classifier. Since a very large set of features (the feature set contained 222 different features), a feature selection method based on greedy forward strategy was used. By taking a ”greedy approach”, the feature that produces the highest score when combined with an already selected set of features is chosen. The selection ends when no new feature increases the classifier performance. In this research, we used the feature set which was selected when training on the i2b2 dataset, as we did not produce any

[2] UMLS is a database of medical terms and the relationships between them sponsored by the NIH.

additional annotations for medical concepts alone. The feature set was reported in [16].

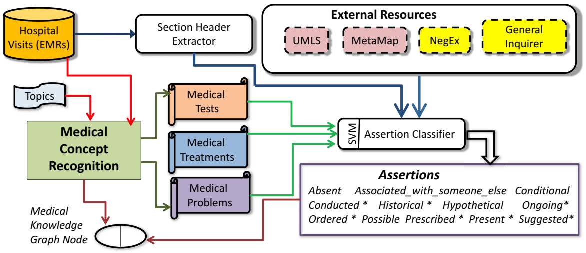

Fig. 4: Assertion classification

B. Assigning Belief Values to Medical Concepts

To be able to encode the medical knowledge in the medical knowledge graph (QMKG), we also needed to automatically identify whether a medical concept is qualified by any of the assertions, given in Table II. To be able to make such assertions, we cast this problem as a classification problem, implemented as an SVM classifier which is influenced by (a) the medical concepts on which the assertion is produced, (b) the meta data available in the section header where the assertion is implied and (c) features available from UMLS (extracted by MetaMap) as well as features reflective of negated statements, disclosed through the NegEx negation detection package. A special case of features that provide belief values are available from the General Inquirer's category information. The General Inquirer [19] is the first general-purpose computerized text analysis resource developed in psychology, relying on the Harvard psychological dictionaries. The current version of the Harvard General Inquirer encodes several hundreds of semantic categories along with lexical categories that correlate with mental states, motives, social and cultural notes, as well as different aspects of general distress. As in [16], we have used only the IF category from the Inquirer, as it encodes uncertainty. For a belief value, such as ABSENT, the capability of detecting the negation word associated with the medical concept was important, thus NegEx provided important information. For the detection of assertions, we performed additional annotations on the TRECMed clinical data, to provide training data for six additional values which are marked with an '*' in Figure 4. The assertion classifier was re-trained and the same 27 features reported in the state-of-the-art assertion identification method from [16] were used. As Figure 4 illustrates, each node from the QMKG is generated when the medical concept recognition and the assertion classifiers are used both on the topics and in the EMRs. To generate the edges of the graph and to learn their strengths, the methods detailed in Section 3 were employed. The QMKG is used to enhance the results of our patient cohort retrieval system, which is described in Section 4.

III. Constructing the Medical Knowledge Graph

An intuitive means of constructing a medical knowledge graph is to create a node for each encountered medical concept and associated assertion value in the corpus (henceforth referred to as a "grounded concept"), and an edge $e = (u, v)$ between grounded concepts u and v if and only if they co-occur within the same context. In this paper, we consider a context as a window of 20 tokens. Let $\lambda(i)$ represent the i-th document context for $0 \leq i \leq M$. More formally, given a sequence of context windows, $w_0, w_1, \ldots, w_N$, we can construct an adjacency matrix $A^{\lambda(i)} \in \mathbb{R}^{|V| \times |V|}$ where $|V|$ is total number of observed grounded concepts, and each element within $A^{\lambda(i)}$ is defined as

$$A_{u,v}^{\lambda(k)} = A^{\lambda(k)} + \begin{cases} 1 & \text{if grounded concept } u \text{ co-occurs} \\ & \text{with grounded concept } v \text{ within} \\ & \text{context } \lambda(i) \\ 0 & \text{otherwise.} \end{cases}$$

The u-th row vector $V_{u,\star}^{\lambda(i)}$ of $V^{\lambda(i)}$ corresponds to the number of times grounded-concept v has co-occurred with u within i contexts. Likewise, the v-th column vector $V_{\star,v}^{\lambda(i)}$ of $V^{\lambda(i)}$ denotes number of co-occurrences between u and v within the first i contexts. All non-zero entries in A correspond to all edges of the medical knowledge graph $G = (V, E)$, where V is the set of vertices corresponding to all grounded concepts, and E is the set of edges corresponding to all non-zero entries in A^M such that edge $e = (u, v) \in E$ if and only if $A_{u,v}^M \neq 0$.

Likewise, we encode the similarity between two grounded concepts within the matrix $W_{|V|^2} \in \mathbb{R}^{|V| \times |V|}$, where each element within W is defined as

$$W_{u,v} = S(u, v)$$

where S is some similarity function $S : (V \times V) \to \mathbb{R}$ and V is the set of possible vertices (grounded concepts) in G.

A. First Order Similarity

The weights of the QMKG, W, correspond to semantic similarity scores between two grounded medical concepts. Although, there are a variety of techniques for calculating the semantic similarity between two spans of text, we consider four such techniques.

$\boxed{1}$ The first technique we used for qualifying the similarity between two qualified medical concepts is using the point-wise mutual-information (PMI) between two qualified medical concepts. To compute the PMI between u and v, we make use of equation 1, given in Table III. For example, the PMI between *heart failure*/PRESENT and *new cardiac event*/PRESENT is 15.48. Likewise, the PMI between *progressive exertional shortness of breath*/PRESENT is 15.26. These PMI values indicate that the independence between these pairs of qualified medical concepts is low, and, thus, that they are likely to be related.

$\boxed{2}$ However, point-wise mutual information has a well-known bias towards infrequent events. This is clear when one attempts to extrapolate knowledge from the top-scoring PMI-weighted edges for a given grounded concept. Consider, that for the PMI between *heart failure*/PRESENT and the unrelated qualified medical concept *a divert colostomy*/CONDITIONAL is 15.08. Equation 2, often referred to as Lin's modified

$$\mathrm{pmi}(u,v) = \log \frac{A^M(u,v)}{A^M(u,\star) * A^M(\star,v)} \tag{1}$$

$$\mathrm{lin\_pmi}(u,v) = \frac{A^M(u,v)}{A^M(u,v)+1} \times \frac{\min\left(A^M(u,\star), A^M(\star,v)\right)}{\min\left(A^M(u,\star)A^M(\star,v)\right)+1} \times \mathrm{pmi}(u,v) \tag{2}$$

$$\mathrm{log\_fisher}(u,v) = \log\left[\frac{C\left(A^M(u,\star), A^M(u,v)\right) \times C\left(|V| - A^M(u,\star), A^M(\star,v) - A^M(u,v)\right)}{C\left(|V|, A^M(u,\star) + A^M(\star,v)\right)}\right] \tag{3}$$

$$\mathrm{ngd}(u,v) = \frac{\log \big(\max \big[A^M(u,\star), A^M(\star,v)\big]\big] - \log\big[A^M(u,v)\big]}{\log|V| - \log\left[\min\left(A^M(u,\star), A^M(\star,v)\right)\right]} \tag{4}$$

TABLE III: Equations used for computing the first order similarity between qualified medical concepts. Where $C(n,k)$ denotes the binomial coefficient, n choose k; $A^M(\star,v) = \sum_i A^M(i,v)$; $A^M(u,\star) = \sum_j A^M(u,j)$; and $|V|$ is the vocabulary size – the number of qualified medical concepts.

PMI, addresses this bias by scaling the PMI by a discounting factor given in [20]. This discounting factor considers the frequency of each individual qualified medical concept in a way that discourages extremely rare qualified medical concepts from having too much impact on the resulting weight. For comparison, the highest scoring edges using Lin's modified PMI for *heart failure*/PRESENT are *nonischemic cardiomyopathy*/PRESENT, and *intravaneous lasix*/ONGOING. Both are commonly associated with heart failure. Additionally, the motivating example – *a divert colostomy*/CONDITIONAL – given above has a much lower relative weight.

$\boxed{3}$ We investigate Fisher's exact test which measures the significance of association (contingency) between two vertices in the graph, and given in Equation 3. Fisher's exact test is commonly used in statistics to evaluate the null hypothesis in situations where the sample size is too small to evaluate using the Chi-squared test. Note that fisher's exact test measures the difference between the proportions of two events, and thus when used to measure similarity, the least weight edges are the most similar. Continuing our example, the most similar neighbors for *heart failure*/PRESENT according to Fisher's exact test are *hypertension*/PRESENT at -116.57, and *congestive heart failure*/PRESENT at -92.3.

$\boxed{4}$ Our fourth technique, equation 4, adapts the Normalized Google Distance [21], which is a way of measuring semantic similarity based on Google hits, into a similarity measure for grounded medical concepts. We do this by replacing the Google frequency with the number of associated contexts in our corpus. Again, like with Fisher's exact test, the smallest weighted edges are the most similar. As such, *our complete left bundle branch block*/PRESENT and *nonischemic cardiomyopathy*/PRESENT are the two most similar neighbors for *heart failure*/PRESENT.

B. Second Order Similarity

Because of the incredible sparsity of qualified medical concepts in EMRs, there are a multitude of qualified medical concepts that do not share the same context window, but still share semantic similarity that would be of value to medical knowledge processing systems. For example, consider the concepts *atrial fibrillation* and *ventricular fibrillation*. These concepts are unlikely to co-occur directly, however, they represent the same medical phenomena: (irregular heart beat) but occur in different anatomical locations of the heart: the atrium, and the ventricles, respectively. In order to capture this relationship, we generalize the notion of second-order PMI which has been exploited for learning synonymy, to build a measure of second-order similarity given any first-

Algorithm 1 Computing the second-order similarity matrix given a graph and its first-order similarity matrix

```
Require: $G = (V, E)$ is a graph of qualified medical concepts
Require: $W$ is a first-order similarity matrix of size $|V| \times |V|$

function SECOND-ORDER-SIM($G = (V,E)$, $W_{|V|^2}$)
    initialize $Z$ as a $|V| \times |V|$ matrix
    for all $e = (u,v) \in E$ do
        $\beta_u \leftarrow \mathrm{floor}\left(\log \sum_i W(u,i)^2 \times \log_2 |V| \div \delta\right)$
        $\beta_v \leftarrow \mathrm{floor}\left(\log \sum_j W(j,v)^2 \times \log_2 |V| \div \delta\right)$
        $z_u \leftarrow$ SUM-TOP-$\beta$($u$, $\beta_u$, $W$)
        $z_v \leftarrow$ SUM-TOP-$\beta$($v$, $\beta_v$, $W$)
        $Z(u,v) \leftarrow z_u {\beta_u}^{-1} + z_v {\beta_v}^{-1}$
    end for
    return $Z$
end function

function SUM-TOP-$\beta$($v$, $\beta$, $W_{V^2}$)
    initialize $Y$ as a zero-vector of size $|V|$
    for $i = 1$ to $|V|$ do
        $Y[i] \leftarrow \sum_j W(j,v)$
    end for
    sort $Y$ in descending order
    $z \leftarrow 0$
    for $i = 1$ to $\beta$ do
        $z \leftarrow z + Y[i]^\gamma$
    end for
    return $z$
end function
```

order similarity function. We provide an algorithm, which we call Second-Order-Sim, for calculating the second-order similarity based on any first-order similarity matrix W for a graph $G = (V, E)$.

The second-order similarity can be viewed as an aggregation of the weights (first-order similarities) on paths connecting any pairs of nodes. In this work, we only consider paths containing a single intermediate node (e.g. $u \leftarrow t \leftarrow$). In order to compute the second order similarity between a pair of nodes (u, v) from the graph, we first need to determine the number of single-intermediate-node paths we will consider. In our case, u and v encode triplets containing the lexicalised medical concept, the concept type, and the assertion. Hence, we want to determine how many intermediary medical concepts should be considered when determining the second-order similarity between u and v. We call these numbers β_u and β_v. The second-order similarity of u to v is then computed based on the first-order similarities along the most similar β_u paths from u to v (and vice-versa for the similarity from v to u). In calculating the values of β_u and β_v, we determine (a) how many other medical concepts encoded in the QMKG may be used to semantically describe the concepts from u and v, and (b) how many nodes should be considered when generating paths between u and v in the QMKG. The algorithm above gives the details about how β_u and β_v are are computed, which enables the estimation of the second-order similarity for nodes u and v, denoted as z_u and z_v. The function $SUM - TOP$ enables the computation of these values. Further details of the motivations behind and computation of the β values are provided in [22]. Finally, we compute the second-order similarity between v and u as a normalized sum of the first-order-similarities in the top β_v and β_u paths between u and v (the normalized sum of z_u and z_v). This second-order similarity encodes the indirect similarity between v and u given β intermediate nodes in the QMKG. That is, if the sum of the top β weights between v and u is significantly large, then the second order similarity between v and u will also be large, indicating that v and u are highly similar.

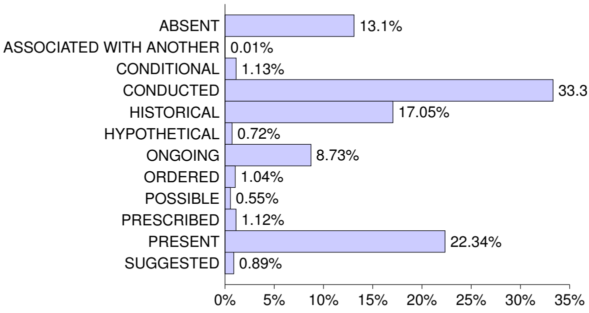

Fig. 5: Distribution of assertions in our collection of EMRS.

IV. Evaluation and Discussion

The QMKG that we have automatically generated contains 634 thousand nodes and 13.9 billion edges (3.45% of nodes are connected), with 53.04% of nodes encoding qualified medical problems, 23.62% of nodes encoding medical tests, and 23.34% of nodes encoding medical treatments. Since the QMKG is unique in its representation of qualified medical concepts, we were also interested in the distribution of the assertions in the QMKG. Figure 5 illustrates the distribution, indicating that CONDUCTED medical tests constitute the majority of the nodes, followed by PRESENT medical problems, and HISTORICAL medical problems, treatments, or tests.

	i2b2 2010			EMRs		
Class	P	R	F_1	P	R	F_1
ABSENT	95.93	93.41	94.65	88.9	89.1	89.0
ASSOCIATED WITH ANOTHER	91.47	81.38	86.13	0	0	-
CONDITIONAL	72.86	29.82	42.32	20.0	63.6	30.4
CONDUCTED				89.8	80.7	85.0
HISTORICAL				57.0	65.5	61.0
HYPOTHETICAL	92.2	87.0	89.5	0	0	-
ONGOING				81.3	64.2	71.7
ORDERED				17.6	54.6	26.7
POSSIBLE	81.63	58.89	68.42	4.3	25.0	7.3
PRESCRIBED				13.6	27.6	18.2
PRESENT	94.39	98.00	96.17	90.1	78.4	83.9
SUGGESTED				0	0	-

TABLE IV: Precision and recall for assertion types as evaluated against the 2010 i2b2 data, and our annotations for EMRs. P denotes the precision, R denotes the recall, and F_1 denotes the F_1 score.

Because the QMKG is automatically generated, we were interested to evaluate the correctness of the encoded information. The medical information from the nodes of the QMKG uses extensions of the techniques reported in [16]. The ability to detect lexicalized medical concepts on the 2010 i2b2 data had an F_1-score of 79.59%. When the i2b2 assertions were used, the system's ability to identify them obtained an F_1-score of 92.75%. The precision and recall varied across classes, as illustrated in Table IV. Clearly, our assertion classification methodology performs best on the classes that occur the most (CONDUCTED, PRESENT, ABSENT), while rarer classes (SUGGESTED, and ORDERED) are harder to detect.

As we did not have the same amount of annotations as those used in the 2010 i2b2 challenge (where 25 thousand medical concepts were used for training and 40 thousand for testing), we have relied on 2,349 new annotations for our new assertions classes that we have introduced. The distribution of the assertions far from uniform within the i2b2 dataset, and so it was in our dataset as well. Evaluating the assertions was was performed in two phases, first, we evaluated the assertions that were used in the i2b2 challenge and then we evaluated the new assertions against our annotations and achieved an accuracy of 75.99%. Table IV illustrates these results. Note that certain assertion values (ASSOCIATED WITH ANOTHER, HYPOTHETICAL, and SUGGESTED) were encountered very rarely (< 10 occurrences) and were not correctly classified. We conclude that our automatic generation of the QMKG achieves results comparable to state-of-the-art techniques.

Because our QMKG was generated with the purpose of improving a patient cohort retrieval system, we have also evaluated the results it enabled on our TRECMed system. For an in-depth discussion, please refer to [5]. A brief overview

Measure	iAP	iNDCG	P @ 10
None*	.2795	.3938	.1447
Lin's PMI	.3256	.5898	.2085
PMI	.3458	.6124	.2319
Fisher's Exact Test	.3286	.5942	.2149
NGD	.3403	.6093	.2319

TABLE V: Scores achieved on the TRECMed 2012 topics. iAP refers to the inferred Average Precision, $iNDCG$ refers to the inferred Normalized Discounted Cumulative Gain, and $P@10$ refers to the precision within the first 10 results.

of our system follows: given a query, keywords are extracted using Wikipedia article titles, and their assertion values are identified. From there, keywords are expanded by selecting the 20 highest-weighted (most similar) neighbors connected to each pair of (keyword, assertion) previously extracted. An initial set of documents are then retrieved from the EMRs and scored according to a weighted BM25 query containing all extracted (keyword, assertion) pairs, weighted by the strength of their relevant incident edge. Finally, documents are re-ranked to ensure that keyword mentions have the desired assertion, and some additional topic constraints (such as the patient's age and gender) are evaluated. Table V displays these results.

Clearly, incorporating the knowledge encoded within the QMKG yielded significantly more relevant patient cohorts. Additionally, and somewhat unexpectedly, the basic PMI similarity measure yield the best performance. This is likely due to the relative rarity of many medical conditions, which – particularly within the medical domain – may have an increased significance due to their perceived severity.

V. Conclusions

An extraordinary breadth of electronic medical records are used throughout the world. These documents contain detailed narratives of the circumstances surrounding a patients treatment, such as surgical reports, patient histories, discourses with the physician, or discharge summaries. Despite the incredible depth of knowledge encoded in these records, they are not readily usable for machine consumption. This is due to the fact that physicians do not state their reasoning behind certain actions, assuming that readers of their records have the domain knowledge required to infer their motivations. This kind of reasoning, however, can be The 2011 and 2012 Text REtrieval conference evaluated the task of retrieving patient cohorts from Electronic Medical Records (EMRs). We showed that by constructing a graph of medical concepts – medical problems, treatments, or tests – qualified by the physician's belief (such as absent, present, or conditional) can greatly improve the relevance of patient cohorts when used for query expansion. Further, we evaluated four different methods for determining the first-order similarity between qualified medical concepts. We also provided a generalized technique for computing the second-order similarity from a first-order similarity matrix. This kind of knowledge – the nature of medical concepts such as problems, treatments, or tests as well as their belief state (e.g. present, hypothetical) – constitutes a reasonable method for systems operating in the medical domain to simulate the high degree of domain knowledge required to interpret the findings in EMRs.

References

[1] C. on Comparative Effective Research Prioritization and I. of Medicine (US), *Initial national priorities for comparative effectiveness research*. National Academies Press, 2009.

[2] E. Voorhees and R. Tong, "Overview of the trec 2011 medical records track," in *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*. Gaithersburg, MD: National Institute for Standards and Technology, 2011.

[3] E. Voorhees and W. Hersh, "Overview of the trec 2012 medical records track," in *The Twenty-First Text REtrieval Conference Proceedings (TREC 2012)*. Gaithersburg, MD: National Institute for Standards and Technology, 2012, unpublished. Draft available at http://trec.nist.gov/.

[4] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall, "2010 i2b2/va challenge on concepts, assertions, and relations in clinical text," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 552–556, 2011.

[5] T. Goodwin, B. Rink, K. Roberts, and S. M. Harabagiu, "Cohort shepherd: Discovering cohort traits from hospital visits," in *Proceedings of The 20th Text REtrieval Conference*, 2011.

[6] T. Goodwin, K. Roberts, B. Rink, and S. M. Harabagiu, "Cohort shepherd ii: Verifying cohort constraints from hospital visits," 2012.

[7] C. Fellbaum, *WordNet: An Electronic Lexical Database*. The MIT press, 1998.

[8] E. M. Voorhees, "Query expansion using lexical-semantic relations," in *SIGIR*. Springer, 1994, pp. 61–69.

[9] E. M. Voorhees and D. Harman, "Overview of the sixth text retrieval conference (trec-6)," in *TREC*, 1997, pp. 1–24.

[10] G. Salton, "The smart retrieval systemexperiments in automatic document processing," 1971.

[11] E. M. Voorhees, "Using wordnet to disambiguate word senses for text retrieval," in *SIGIR*, R. Korfhage, E. M. Rasmussen, and P. Willett, Eds. ACM, 1993, pp. 171–180.

[12] P. Sondhi, J. Sun, H. Tong, and C. Zhai, "Sympgraph: a framework for mining clinical notes through symptom relation graphs," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 1167–1175. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339712

[13] P. Schuyler, W. Hole, M. Tuttle, and D. Sherertz, "The umls metathesaurus: Representing different views of biomedical concepts." *Bulletin of the Medical Library Association*, vol. 81, no. 2, p. 217, 1993.

[14] A. R. Aronson, "Effective mapping of biomedical text to the umls metathesaurus: the metamap program." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 17.

[15] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute, "Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[16] K. Roberts and S. Harabagiu, "A flexible framework for deriving assertions from electronic medical records," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 568–573, 2011.

[17] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl 1, p. D267, 2004.

[18] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii, "Genia corpusa semantically annotated corpus for bio-textmining," *Bioinformatics*, vol. 19, no. suppl 1, pp. i180–i182, 2003.

[19] P. J. Stone, D. C. Dunphy, and M. S. Smith, "The general inquirer: A computer approach to content analysis." 1966.

[20] P. Pantel and D. Lin, "Discovering word senses from text," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 613–619.

[21] R. Cilibrasi and P. M. B. Vitányi, "The google similarity distance," *CoRR*, vol. abs/cs/0412098, 2004.

[22] A. Islam and D. Inkpen, "Second order co-occurrence pmi for determining the semantic similarity of words," in *Proceedings of the International Conference on Language Resources and Evaluation, Genoa, Italy*, 2006, pp. 1033–1038.