Medical Question Answering for Clinical Decision Support

Travis R. Goodwin and Sanda M. Harabagiu
Human Language Technology Research Institute
Department of Computer Science
University of Texas at Dallas
800 W. Campbell Rd.
Richardson, Texas 75080
{travis, sanda}@hlt.utdallas.edu

ABSTRACT

The goal of modern Clinical Decision Support (CDS) systems is to provide physicians with information relevant to their management of patient care. When faced with a medical case, a physician asks questions about the diagnosis, the tests, or treatments that should be administered. Recently, the TREC-CDS track has addressed this challenge by evaluating results of retrieving relevant scientific articles where the answers of medical questions in support of CDS can be found. Although retrieving relevant medical articles instead of identifying the answers was believed to be an easier task, state-of-the-art results are not yet sufficiently promising. In this paper, we present a novel framework for answering medical questions in the spirit of TREC-CDS by first discovering the answer and then selecting and ranking scientific articles that contain the answer. Answer discovery is the result of probabilistic inference which operates on a probabilistic knowledge graph, automatically generated by processing the medical language of large collections of electronic medical records (EMRs). The probabilistic inference of answers combines knowledge from medical practice (EMRs) with knowledge from medical research (scientific articles). It also takes into account the medical knowledge automatically discerned from the medical case description. We show that this novel form of medical question answering (Q/A) produces very promising results in (a) identifying accurately the answers and (b) it improves medical article ranking by 40%.

Keywords

Question Answering; Medical Information Retrieval; Clinical Decision Support

1. INTRODUCTION

In their everyday practice, physicians make a variety of clinical decisions regarding the care of their patients, e.g. deciding the diagnosis, the test(s) or the treatment that they prescribe. Clinical Decision Support (CDS) systems have been designed to help physicians address the myriad of complex clinical decisions that might arise during a patient's care [11]. By leveraging the fact that patient care is documented in electronic medical records (EMRs), one of the goals of modern CDS systems is to anticipate the needs of physicians by linking EMRs with information relevant for patient care. Such relevant information can be retrieved from bio-medical literature. Recently, the special track on Clinical Decision Support in the Text REtrieval Conference (TREC-CDS) [25], has addressed the challenge of retrieving bio-medical articles relevant for to a medical case when answering one of three generic medical questions: (a) "What is the diagnosis?"; (b) "What test(s) should be ordered?"; and (c) "Which treatment(s) should be administered?". The TREC-CDS track did not rely on a collection of EMRs, instead it used an idealized representation of medical records in the form of 30 short medical case reports, each describing a challenging medical case. Thus, systems developed for the TREC-CDS challenge were provided with a list of *topics*, consisting of (1) a narrative describing the fragments from the patient's EMRs that were pertinent to the case; (2) a summary of the medical case and (3) a generic medical question. Systems were expected to use either the medical case description or the summary to answer the question by providing a ranked list of articles available from PubMed Central [30] containing the answers. As only one of the three generic questions was asked in each topic, the expected medical answer type (EMAT) of the question was diagnosis, test or treatment. Figure 1 illustrates three examples of topics evaluated in the 2015 TREC CDS, one example per EMAT. Figure 1 also illustrates the correct answer of each of the questions.

In the 2015 TREC-CDS track a new task was offered, in which for questions having the EMAT $\in$ {test,treatment}, the patient's diagnosis was provided (shown in Figure 1). The results for this new task, as reported in [25] were superior to the results for the same topics when no diagnoses were provided. This observation let us to believe that when knowing even a partial answer to the question, the ability to retrieve relevant bio-medical literature was significantly improved. Moreover, we asked ourselves if identifying the answers to the medical questions could be performed with acceptable accuracy. More importantly, we wondered if we should first try to find the answer and then rank the relevant scientific articles for a given question. It was clear to us from the beginning that answer identification would be a harder problem, unless we could tap into a new form of knowledge and consider answering the questions directly from a knowledge base (KB). Question answering (Q/A) from KBs has experienced a recent revival. In the 60's and 70's, domain-specific knowledge bases were used to support Q/A, e.g. the Lunar Q/A system [33]. With the recent growth of KBs such as DBPedia [3] and Freebase [7], new promising methods for Q/A from KBs have emerged [9, 34, 5]. These methods map questions into sophisticated meaning-representations which are used to retrieve the answers from the KB.

In this paper, we present a novel Q/A from KB method used for answering the medical questions evaluated in TREC-CDS. First, instead of relying on an existing large KB, we automatically generated a very large medical knowledge graph from a publicly available collection of EMRs. As reported in [22], the medical

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983819

Topic 32	Topic 44	Topic 53
EMAT: Diagnosis **Description:** A 44-year-old man was recently in an automobile accident where he sustained a skull fracture. In the emergency room, he noted clear fluid dripping from his nose. The following day he started complaining of severe headache and fever. Nuchal rigidity was found on physical examination. **Summary:** A 44-year-old man with coffee-ground emesis, tachycardia, hypoxia, hypotension and cool, clammy extremities. **Answer:** Spinal Tap; Cerebrospinal Fluid (CSF) Analysis	**EMAT:** Test **Description:** A 27-year-old woman at 11 weeks gestation in her second pregnancy is found to have a hemoglobin (Hb) of 9.0 g/dL, white blood cell count 6.3×109/L, platelet count 119×109/L, mean corpuscular volume 109 fL. [...] Negative DAT, normal clotting screen, elevated LDH (2000 IU/L), normal urea and electrolytes, normal alanine aminotransferase (ALT), anisocytosis, poikilocytosis, no fragments, no agglutination, polychromasia and presence of hemosiderin in the urine. **Summary:** A young woman in her second gestation presenting with anemia resistant to improvement by iron supplementation, elevated LDH, anisocytosis, poikilocytosis, hemosiderinuria and normal clotting screen. **Diagnosis:** Paroxysmal Nocturnal Hemoglobinuria **Answer:** Flow cytometry for Glypiated (GPI-linked) Proteins	**EMAT:** Treatment **Description:** An 18-year-old male returning from a recent vacation in Asia presents to the ER with a sudden onset of high fever, chills, facial flushing, epistaxis and severe headache and joint pain. His complete blood count reveals leukopenia, increased hematocrit concentration and thrombocytopenia. **Summary:** An 18-year-old male returned from Asia a week ago. He presents with high fever, severe headache and joint pain. His blood analysis reveals leukopenia, increased hematocrit and thrombocytopenia. **Diagnosis:** Dengue fever **Answer:** Supportive Care with Analgesics; Careful Fluid Management

Figure 1: Examples of topics evaluated in the 2015 TREC CDS track.

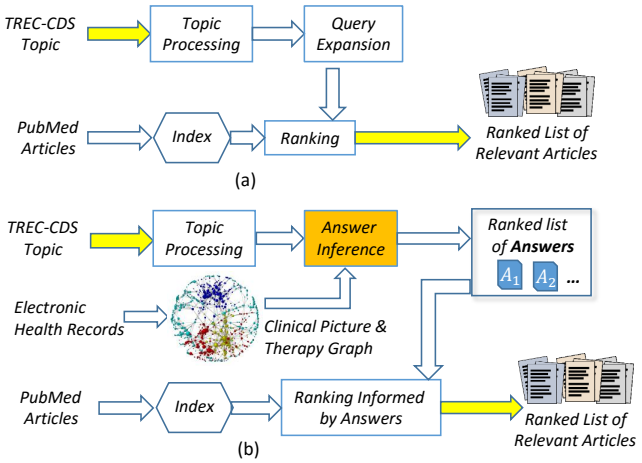

Figure 2: Architectures of medical question answering systems for clinical decision support

case descriptions from the TREC-CDS topics were generated by consulting the EMRs from MIMIC-II [16]. Consequently, we used all the publicly available EMRs provided by MIMIC-III (a more recent superset of the EMRs in MIMIC-II) to automatically generate a very large probabilistic knowledge graph designed to encode knowledge acquired from medical practice. Second, instead of retrieving answers directly from the KB, the answers were obtained through probabilistic inference methods. Third, instead of identifying the answers from relevant PubMed articles, we used the answers inferred from the knowledge graph to select and rank the PubMed articles which contain them. In this way, we replaced the architecture of the typical system that was evaluated in TREC-CDS, illustrated in Figure 2(a) with a new architecture, illustrated in Figure 2(b). The new architecture identifies the ranked list of answers to the questions as well as the ranked list of scientific articles that contain them. While current state-of-the-art systems that were evaluated in TREC-CDS first processed the topics and then used query expansion methods to enhance the relevance of the retrieved scientific articles from PubMed, as reported in [22] and illustrated in Figure 2(a), we relied on a knowledge graph encoding the clinical picture and therapy of a large population of patients documented in the MIMIC III EMR database. The automatic generation of the knowledge graph involved: (1) medical language processing to identify medical concepts representing signs/symptoms, diagnoses, tests and treatments, as well as their assertions; (2) the cohesive properties of the medical narratives from the EMRs; and (3) a factorized Markov network representation of the medical knowledge. As illustrated in Figure 2(b), this probabilistic knowledge graph was used for inferring the answer of the TREC-CDS topics, which were processed to discern the medical concepts and their assertions in the same format as the nodes from the knowledge graph. We experimented with three different probabilistic inference methods to identify the most likely answers for each of the TREC-CDS topics evaluated in 2015. By participating in the challenge, we had access to the correct answers [1], thus we could evaluate the correctness of the answers identified by our novel Q/A from KB method. Moreover, the inferred answers allowed us to produce a ranking of the scientific articles that contained them and thus define a novel, answer-informed relevance model. Our main contributions in this paper are:

1. Answering medical questions related to complex medical cases from an automatically generated knowledge base derived from a vast, publicly available EMR collection;
2. Using probabilistic inference to identify answers from a vast medical knowledge graph;
3. Combining medical knowledge derived from an EMR collection with medical knowledge derived from relevant scientific articles to enhance the quality of probabilistic inference of medical answers – a combination that unifies medical knowledge characterizing medical practice (from EMRs) with medical knowledge characterizing medical research (from scientific articles); and
4. Using the likelihood of the automatically discovered answers to the question associated with a topic to produce a novel ranking of the relevant scientific articles containing the answers.

The remainder of the paper is organized as follows. Section 2 details the answer inference in the new architecture for Q/A-CDS. Section 3 details the automatic generation of the medical knowledge graph while Section 4 presents the three forms of probabilistic inference of answers we experimented with. Section 5 discusses the experimental results and Section 6 summarizes the conclusions.

2. AN ARCHITECTURE FOR INFERRING MEDICAL ANSWERS

The cornerstone of our medical Q/A method for clinical decision support (CDS) is the derivation of the answers to a topic's question from a vast medical knowledge graph, generated automatically from a collection of EMRs. The medical knowledge base contained approximately 634 thousand nodes and 14 billion edges, in which each node represents a medical concepts and the belief value (or assertion, as described in Section 3.3) associated with

[1]All participants in the 2015 TREC-CDS track were provided with the correct answers to the topics a few months after the evaluation.

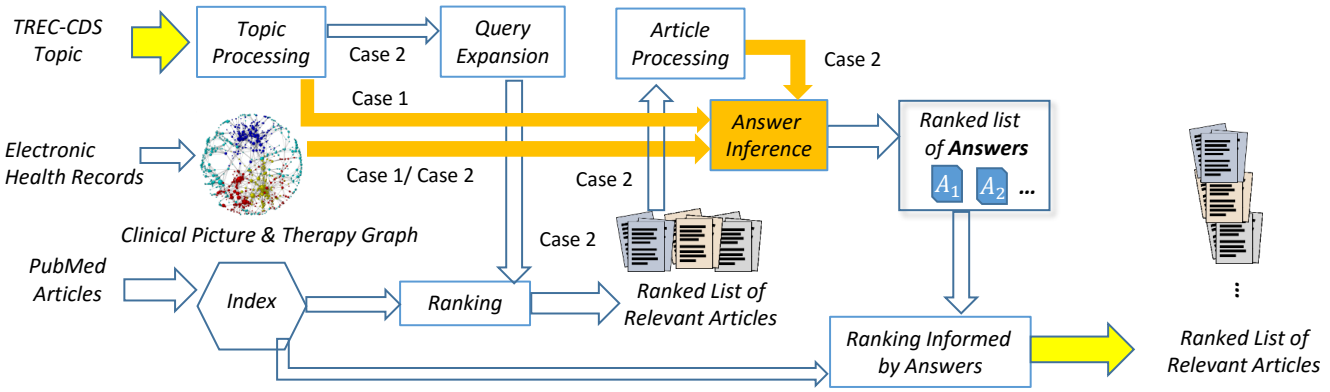

Figure 3: An architecture that implements two different cases for answering medical questions for clinical decision support

it. We automatically identified four types of medical concepts: signs/symptoms, tests, diagnoses and treatments (as detailed in Section 3.2). However, identifying medical concepts is not sufficient to capture all the subtleties of medical language used by physicians when expressing medical knowledge. Medical science involves asking hypotheses, experimenting with treatments, and formulating beliefs about the diagnoses and tests. Therefore, when writing about medical concepts, physicians often use hedging as a linguistic means of expressing an opinion rather than a fact. Consequently, clinical writing reflects this *modus operandi* with a rich set of speculative statements. Hence, automatically discovering clinical knowledge from EMRs needs to take into account the physician's degree of belief by qualifying the medical concepts with assertions indicating the physician's belief value (e.g. HYPOTHETICAL, PRESENT, ABSENT) as detailed in Section 3.3. It should be noted that the same medical language processing techniques were used to process the EMRs, the medical topics, and the (relevant) scientific articles from PubMed.

To represent the relations spanning the medical concepts in our knowledge graph, we modeled the cohesive properties of the narratives from EMRs (details are provided in Section 3.1). In order to use this graph to infer answers to medical questions we cast the medical knowledge graph as a factorized Markov network, which is a type probabilistic graphical model. We call this graphical model the *clinical picture and therapy graph* (CPTG) because it enables us to compute the probability distribution over *all the possible clinical pictures and therapies of patients*. For a given topic t, the set of medical concepts and their assertions discerned from t is interpreted as a *sketch* of the clinical picture and therapy described in the topic, represented as $Z(t)$. Thus, answering the medical question associated with t amounts to determining which medical concept in the CPTG (with the same type as the expected medical answer type or EMAT) has the highest likelihood given $Z(t)$. In addition to the medical sketch $Z(t)$, we believed that answering medical questions with the CPTG could benefit from combining the medical sketch of a topic with knowledge acquired from individual scientific articles deemed relevant to the topic. This belief is also motivated by our observation that the joint distribution represented by the CPTG favors more common concepts, whereas the topics evaluated in the TREC-CDS correspond to complex medical cases, rather than common cases. Thus, we believe that the combination of $Z(t)$ with all the medical concepts (and their assertions) derived from a scientific article l relevant to the topic generates a more complete view of a possible clinical picture and therapy for a patient than the one discerned only from the topic. Therefore, we consider this extended set of medical concepts and assertions for a topic t as its *extended sketch*, denoted as $EZ(t,l)$. Regardless of which sketch $z \in \{Z(t), EZ(t,l)\}$ of a topic is used, we discovered the most likely answer $\hat{a}$ to the medical question associated with t by discovering the medical concept which, when combined with the sketch, produces the most likely clinical picture and therapy. Formally:

$$\hat{a} = \underset{a \in A}{\text{argmax}} P(\{a\} \cup z | z) = \frac{P(\{a\} \cup z)}{P(z)} \quad (1)$$

where the set A denotes all the concepts in the CPTG with the same type as the EMAT, and $P(\cdot)$ refers to the probability estimate provided by the CPTG.

The architecture of the medical QA system which operated on the medical graph that we derived automatically from a collection of EMRs (i.e. the CPTG) is represented in Figure 3. When the $Z(t)$ is used in the inference of answers, we have Case 1, as illustrated in Figure 3, in which the topic is processed to derive $Z(t)$ by automatically identifying medical concepts and assertions from either the description or the summary of the topic. When $EZ(t,l)$ (the extended sketch of the topic) is used, in addition to the concepts (and their assertions) obtained from processing the topic, new knowledge from a relevant scientific article l needs to be derived. It is to be noted that from each relevant article for a topic t we generated a distinct extended sketch which combines $Z(t)$ with the medical concepts (and their assertions) discovered in the article. In Figure 3, this is represented as Case 2, in which the first step is to derive (and expand) a query for the topic.

When processing the topic to generate a query, deciding whether to use individual words or concepts is important. In TREC-CDS, there were systems that used all the content words from the description to produce the query, while other systems considered only medical concepts. Both the Unified Medical Language System (UMLS) [6] and MeSH [17] were commonly used as ontological resources for medical concepts. In the architecture represented in Figure 3, we opted to use medical concepts rather than words, identifying the signs, symptoms, diagnoses, tests and treatments mentioned in the topic. We expanded each query with medical concepts from UMLS which share the same concept unique identifier (CUI) as any concept detected from the topic, obtaining synonyms, and, in some cases, hyponyms, and hypernyms. Some medical concepts (e.g. "crystalloid solution") or their synonyms (e.g. "sodium chloride") are phrases rather than

single words. Consequently, the resulting expanded query consists of lists of key-phrases which are processed by a relevance model to retrieve and rank articles from an index of PubMed articles. The index was generated from the PubMed Central Open Access Subset. We used a snapshot of these articles from January 21, 2014 containing a total of 733,138 articles which were provided by the TREC-CDS organizers. In TREC-CDS, systems implemented a variety of relevance models, as reported in [22]; we experimented with several relevance models, discussed in the evaluation section. For each of the first 1,000 relevant articles[2], which we denote as L , we automatically identified the medical concepts and their assertions. This enabled us to generate for each article, $l_r \in L$ its extended medical sketch of the topic t, denoted as $EZ(t,l_r)$.

A close inspection of the contents of the extended medical sketches obtained for many scientific articles indicated the inclusion of many medical concepts which presented no relevance to the topic. This reflects the fact that many of the scientific articles in PubMed Central document unexpected or unusual medical cases – often in non-human subjects. This created a serious problem in the usage of the extended medical sketches to infer answers from the CPTG. Specifically, because the likelihood estimate of an answer enabled by the CPTG is based on the observed clinical pictures and therapies of patients in the MIMIC clinical database, non-relevant scientific articles which contained common diagnoses, treatments, tests, signs, or symptoms had a disproportionately large impact on the ranking of answers. In order to solve this problem, we refined the ranking of answers provided in Equation 1 in order to incorporate the relevance of the scientific article used for creating each extended medical sketch. Thus, in Case 2 illustrated in Figure 3, we produced the answer ranking by using a novel probabilistic metric, namely the Reciprocal-Rank Conditional Score (RRCS). RRCS considers for each article in L, (1) the conditional probability of the answer given the extended sketch associated with that article, i.e. $EZ(t,l_r)$, as well as (2) the relevance rank of the article, represented by the rank r of l_r in L. Formally, the new ranking of answers to a question associated with topic t generated by the RRCS metric is defined as :

$$RRCS(a) = \sum_{r=1}^{|L|} \frac{1}{r} \cdot P(\{a\} \cup EZ(t,l_r) | EZ(t,l_r)) \quad (2)$$

The ranking of answers based on the RRCS reflects both the likelihood of the answer according to $EZ(t,l_r)$ as well as the relevance of each scientific article.

In addition to ranking medical answers, we also use the CPTG to rank scientific articles based on the answers they contain. In Case 1, when the medical sketch of a topic $Z(t)$ was used for inferring the answer, the ranked list of answers was produced by relying entirely on the knowledge from the medical sketch and from the EMR collection. Therefore, the set of scientific articles that contain at least one of the answers of the medical question for a topic t in Case 1 needs to be retrieved. Hence, a query in disjunctive form of all the inferred answers is used. When the relevance model uses the query against the index, it provides a list of ranked relevant articles L. We denote by Y_i all answers found in a relevant article ranked on position i of L. This allows us to define the relevance of scientific articles responding to the answer of a topic t as:

$$\text{Rel}(L_i) = P(Y_i | Z(t)) \propto P(Y_i \cup Z(t)) \quad (3)$$

Equation 3 represents a ranking of each scientific article $l_i \in L$ based on the likelihood of the answers in the article given the medical sketch derived for the topic.

[2]Although we could have considered every possible PubMed article, we limited our experiments to only the top one-thousand articles to improve computational efficiency.

In contrast, in Case 2 represented in Figure 3, when $EZ(t,l)$ is used for inferring the answers, the set of ranked relevant scientific articles L is known, as it was already used to provide the new knowledge for each $EZ(t,l_i)$. Moreover, each answer found in any scientific article from L is necessarily part of its extended medical sketch, thus $Y_i \subseteq EZ(t,l_i)$. Consequently, we define the relevance of scientific articles responding to the answer of a topic t as:

$$\text{Rel}(l_i) = P(Y_i | EZ(t,l_i)) = \frac{P(EZ(t,l_i))}{P(EZ(t,l_i) - Y_i)} \quad (4)$$

In this way the relevance of an article responding to the question of a topic is computed by comparing the likelihood of the extended medical sketch which includes the answers found in the article against the likelihood of the extended medical sketch which does not contain the answers found in the article. Clearly, the ability to infer the probability of a combination of medical concepts (and their assertions), i.e. the capability of determining $P(\cdot)$, enabled us to produce the new, answer-informed ranking models for answers as well as scientific articles given in Equations 1-4.

3. GENERATING THE CLINICAL PICTURE AND THERAPY GRAPH

As defined in [24] the *clinical picture* constitutes the clinical findings about a patient (e.g. medical problems, signs, symptoms, and tests) that might influence the diagnosis. In addition, *therapy* is defined as the set of all treatments, cures, and preventions included within the management plan for a patient. Moreover the clinical picture may vary significantly between patients with the same disease and may even vary between different points in time for the same patient during the course of his/her disease. Therefore, in order to capture the *variation* in the clinical picture and therapy of a *patient population*, we created a clinical picture and therapy graph (CPTG) in which each node corresponds to a medical concept qualified by its assertion. Inspired by the approach reported in [12], we represented the CPTG as a 4-partite graph in which partitions of nodes represent all the signs/symptoms ($\mathbb{S}$), all the diagnoses ($\mathbb{D}$), all the tests ($\mathbb{E}$) and all the treatments ($\mathbb{R}$) which were automatically recognized in the MIMIC-III[3] collection of EMRs. By considering four partitions of medical concepts, we need to encode all six possible types of relations between each partition of nodes in the CPTG. The relations types are: (1) $\mathbb{S} \longleftrightarrow \mathbb{D}$, between signs/symptoms and diagnoses; (2) $\mathbb{S} \longleftrightarrow \mathbb{E}$, between signs/symptoms and tests; (3) $\mathbb{S} \longleftrightarrow \mathbb{R}$, between signs/symptoms and treatments; (4) $\mathbb{D} \longleftrightarrow \mathbb{E}$, between diagnoses and tests; (5) $\mathbb{E} \longleftrightarrow \mathbb{R}$, between tests and treatments; and (6) $\mathbb{D} \longleftrightarrow \mathbb{R}$, between diagnoses and treatments. We take advantage of the fact that any k-partite graph can be interpreted as a factorized Markov network [15], and encode the strength of these relations using mathematical factors. Figure 4 illustrates the factorized Markov network corresponding to the CPTG.

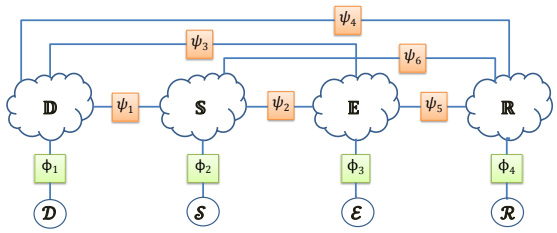

Figure 4: Factorized Markov network modeling the likelihood of any possible clinical picture and therapy.

[3]https://mimic.physionet.org/

A factorized Markov network is a type of Probabilistic Graphical Model which represents knowledge in terms of (1) statistical random variables, and (2) mathematical factors (or functions), which assign a real value to each potential assignment of a set of random variables (known as the factor's scope) allowing us to represent the strength of the relationships between the random variables in the model. In this representation, each possible medical concept (i.e. each node) is interpreted as a binary random variable. This allows any possible clinical picture and therapy (CPT) to be encoded by assigning a value of 1 to the random variable associated with each concept in the CPT which is asserted to be PRESENT, CONDUCTED, ORDERED, or PRESCRIBED, and a value of 0 to the random variable associated with each medical concept which is asserted to be ABSENT. Every other medical concept (i.e. those mentioned with another assertion, or which were not mentioned in the CPT) is considered a *latent* variable whose value is later inferred. This random variable representation allows us to encode **any possible combination of medical concepts (and their assertions)**, which we represent as $\mathcal{C}=\mathcal{D}\cup\mathcal{S}\cup\mathcal{E}\cup\mathcal{R}$, in which $\mathcal{D}\subseteq\mathbb{D}$ represents the random variables corresponding to diagnoses, $\mathcal{S}\subseteq\mathbb{S}$ represents the random variables corresponding to signs/symptoms, $\mathcal{E}\subseteq\mathbb{E}$ represents the random variables corresponding to tests, and $\mathcal{R}\subseteq\mathbb{R}$ represents the random variables corresponding to treatments. We can estimate the likelihood of $\mathcal{C}$ as:

$$MLE(\mathcal{C})=\frac{\text{the number of EMRs with } \mathcal{C} \text{ in the CPT}}{\text{the total number of EMRs}} \quad (5)$$

Using the maximum likelihood estimate provided by $MLE(\mathcal{C})$, we have defined four factors which represent the (prior) probability of a CPT containing combinations of medical concepts with the same type: (1) $\phi_1(\mathcal{D})\propto MLE(\mathcal{D})$, the likelihood of a CPT containing the diagnoses in $\mathcal{D}$; (2) $\phi_2(\mathcal{S})\propto MLE(\mathcal{S})$, the likelihood of a CPT containing the signs/symptoms given by $\mathcal{S}$; (3) $\phi_3(\mathcal{E})\propto MLE(\mathcal{E})$, the likelihood of a CPT containing the tests in $\mathcal{E}$; and (4) $\phi_4(\mathcal{R})\propto MLE(\mathcal{R})$, the likelihood of a CPT containing the treatments in $\mathcal{R}$.

The clinical picture and therapy of a patient also involves relationships between medical concepts of different types. To model these six types of relations, we considered six additional factors: (1) $\psi_1(\mathbb{D},\mathbb{S})\propto MLE(\mathbb{D}\cup\mathbb{S})$, the correlation between all the diagnoses in $\mathbb{D}$ and all the signs/symptoms in $\mathbb{S}$; (2) $\psi_2(\mathbb{S},\mathbb{E})\propto P(\mathbb{S}\cup\mathbb{E})$, the correlation between all the signs/symptoms in $\mathbb{S}$ and all the tests in $\mathbb{E}$; (3) $\psi_3(\mathbb{D},\mathbb{E})\propto P(\mathbb{D}\cup\mathbb{E})$, the correlation between all the diagnoses in $\mathbb{D}$ and all the tests in $\mathbb{E}$; (4) $\psi_4(\mathbb{D},\mathbb{R})\propto P(\mathbb{D}\cup\mathbb{R})$, the correlation between all the diagnoses in $\mathbb{D}$ and all the treatments in $\mathbb{R}$; (5) $\psi_5(\mathbb{E},\mathbb{R})\propto P(\mathbb{E}\cup\mathbb{R})$, the correlation between all the tests in $\mathbb{E}$ and all the treatments in $\mathbb{R}$; and (6) $\psi_6(\mathbb{S},\mathbb{R})\propto P(\mathbb{S}\cup\mathbb{R})$, the correlation between all the signs/symptoms in $\mathbb{S}$ and all the treatments in $\mathbb{R}$.

All ten factors enable us to infer the probability of any possible clinical picture and therapy, such as the sketch $z\in\{Z(t),EZ(t,l)\}$ of the clinical picture and therapy discussed in Section 2. By definition, an (extended) sketch is a set of medical concepts and their assertions. Thus, we can encode z in terms of the random variables corresponding to the diagnoses ($\mathcal{D}$), signs/symptoms ($\mathcal{S}$), tests ($\mathcal{E}$), and treatments $\mathcal{R}$ the (extended) sketch contains, such that $z=\mathcal{D}\cup\mathcal{S}\cup\mathcal{E}\cup\mathcal{R}$. As before, all the remaining random variables (i.e. $\mathbb{D}\cup\mathbb{S}\cup\mathbb{E}\cup\mathbb{R}-z$) are left as latent variables. Using the CPTG (illustrated in Figure 4), we can compute the probability of a medical sketch z:

$$\begin{aligned}P(z)=P(\mathcal{D},\mathcal{S},\mathcal{E},\mathcal{R})\propto&\phi_1(\mathcal{D})\times\phi_2(\mathcal{S})\times\phi_3(\mathcal{E})\times\phi_4(\mathcal{R})\\&\times\psi_1(\mathcal{D},\mathcal{S})\times\psi_2(\mathcal{S},\mathcal{E})\times\psi_3(\mathcal{D},\mathcal{E})\\&\times\psi_4(\mathcal{D},\mathcal{R})\times\psi_5(\mathcal{E},\mathcal{R})\times\psi_6(\mathcal{S},\mathcal{R})\end{aligned} \quad (6)$$

Or, the more compact equivalent notation:

$$P(z)\propto\prod_{i=1}^{6}\psi_i(z)\prod_{j=1}^{4}\phi_j(z) \quad (7)$$

where, for each factor, we ignore all medical concepts which do not have the semantic types expected by that factor's scope (e.g. $\psi_1(z):=\psi_1(\mathcal{D},\mathcal{S})$). As defined, the probability distribution given in Equation 7 is the product of the ten factors defined above. Each of these factors depends on the maximum likelihood estimate of how many CPTs in the EMR collection contain various combinations of medical concepts. Thus, computing $P(z)$ relies on the ability to automatically recognize medical concepts and their assertions in EMRs, as described in Sections 3.1 and 3.2.

3.1 Identification of Medical Concepts

Our methodology for automatically recognizing medical concepts in clinical texts benefits from the general framework developed by the 2010 shared-task on Challenges in Natural Language Processing for Clinical Data [29] provided by the Informatics for Integrating Biology at he Bedside (i2b2) and the United States Department of Veteran's Affairs (VA). In this challenge, identification of medical concepts in clinical narratives targeted three categories: medical problems, treatments, or tests. In this work, we have extended this framework to also distinguish two subtypes of medical problems: (1) signs (observations from a physical exam) and symptoms (observations by the patient); and (2) the diagnoses, including co-morbid diseases or disorders. Figure 6 illustrates our methodology.

We followed the framework reported by [21] in which medical concept identification was cast as a three-stage classification problem, using 72,846 annotations of medical concepts and their assertions provided by the i2b2 challenge. In the first stage, a conditional random field (CRF) is used to determine the boundaries (starting and ending tokens) of each medical concept. In the second stage, a support vector machine (SVM) was used to classify each type of medical concept into a medical problem, treatment, or test.

Classification relied primarily on lexical features, as well as concept type information from UMLS and Wikipedia and predicate-argument semantics resulting from automatic feature selection as described in [21]. The resources used for feature extraction are The Unified Medical Language System (UMLS) [6], MetaMap [2], the GENIA project [13], WordNet [10], PropBank [14], the SwiRL semantic role labeler [28], and Wikipedia.

Unlike the work of [21], we introduced a third and final stage, in which we project each concept onto the UMLS ontology and classify it as a sign or symptom if the UMLS semantic type is SYMPTOM OR SIGN or FINDING, and as a diagnosis otherwise. Finally, synonyms for each medical concept are provided by (1) identifying all UMLS atoms that share the same concept unique identifier (CUI) and (2) groups of article titles in Wikipedia which redirect to the same article. Thus, we can account for synonymous concepts in the CPTG by combining all the nodes corresponding to synonymous concepts. Figure 5 illustrates the results of medical concept recognition.

3.2 Recognizing the Medical Assertions

We followed the framework reported in [21] in which the belief status (or assertion type) of a medical concept is determined by a single SVM classifier. Medical assertions were categorized as PRESENT, ABSENT, POSSIBLE, HYPOTHETICAL, CONDITIONAL, or ASSOCIATED-WITH-SOMEONE-ELSE which were defined in the 2010 i2b2 challenge only for medical problems. Note that we have extended these assertion values to qualify tests and treatments as well as previously reported in [citation withheld]. We considered and annotated five new assertion values to encompass the physicians' beliefs about tests and treatments: PRESCRIBED, ONGOING, and SUGGESTED for treatments as well as ORDERED and CONDUCTED for tests. We have produced an additional set of 2,349 new annotations for the new assertion values as previously reported in [citation withheld].

Description	Diagnoses:	∅
	Signs & Symptoms:	{[**non-productive cough**, dry cough, ...]/PRESENT, [**fever**, febrile, ...]/PRESENT, [**owl's eye inclusion bodies**, owl's eye appearance]/PRESENT, [**infection**, infectious disease, ...]/PRESENT }
	Tests:	{[**bronchoscopy**, tracheobronchial endoscopy, ...]/PRESENT, [**bronchoalveolar lavage**, **BAL**, bronchoalveolar washing, ...]/PRESENT, }
	Treatments:	{[**immunosuppressive drugs**, anti-rejection agent, ...]/PRESENT, [**prednisone**, pregnanes, ...]/PRESENT }
Summmary	Diagnoses:	{[**hypoxia**, hypoxiation, ...]/PRESENT, [**hypotension**, low blood pressure, ...]/PRESENT }
	Signs & Symptoms:	{[**coffee-ground emesis**, coffee ground vomitting, ...]/PRESENT, [**tachycardia**, tachyarrhythmia, ...]/PRESENT, [**clammy extremities**, cold extremities, ...]/PRESENT }
	Tests:	∅
	Treatments:	∅
Article	**Excerpt:**	*...he had high-grade fever (103 °F) with generalized weakness and myalgia. On admission (April 2005), he was febrile (104 °F) with tachycardia (102 beats per minute) and blood pressure of 136/90 mmHg. Systemic examination was normal and there was no hepatomegaly or splenomegaly on per abdominal examination. [...] In view of fever, myalgia, leucopenia, thrombocytopenia, graft dysfunction, and elevated liver enzymes, provisional diagnosis of CMV disease was made ...*
	Diagnoses:	{ [**hepatomegaly**, large liver, ...]/ABSENT, [**splenomegaly**, large spleen, ...]/ABSENT, [**leucopenia**, leucocytopenia, ...]/PRESENT, [**thrombocytopenia**, thrombocytopenic disorder, ...]/PRESENT, [**CMV**, cytomegalovirus, ...]/PRESENT }
	Signs & Symptoms	{ [**fever**, febrile, ...]/PRESENT, [**weakness**, asthenia, ...]/PRESENT, [**myalgia**, muscle pain, ...]/PRESENT, [**tachycardia**, tachyarrhythmia, ...]/PRESENT, [**elevated liver enzymes**, elevated liver enzyme level, ...]PRESENT }
	Tests:	{ [**blood pressure**, BP, ...]/PRESENT, }
	Treatments:	∅

Figure 5: Example of medical concepts and their assertions discerned from the description and summary of medical topic 32 (illustrated in Figure 1) as well as from the relevant PubMed article PMC3132335. Medical concepts mentioned in the text are typeset in boldface, while synonymous concepts are not.

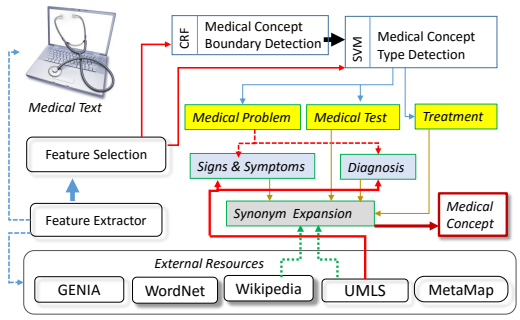

Figure 6: System for Medical Concept Recognition

Our assertion classification methodology relies on the same feature set and external resources reported by [21]: UMLS, MetaMap, NegEx [8] and the Harvard General Inquirer [27]. These external resources, along with lexical features and statistical information about the assertions assigned to previous mentions of the same medical concept in the same document were used to train a multi-class SVM by following the framework reported in [21].

4. ANSWER INFERENCE

Inferring the answers to the question associated with a topic, as defined in Equations 1-4, relies on the ability of the CPTG to model the distribution over all possible CPTs. However, evaluating this distribution(defined in Equation 7) can be prohibitively expensive (it requires storing $2^{|\mathbb{D}|*|\mathbb{S}|*|\mathbb{E}|*|\mathbb{R}|}$ counts). Another problem stems from the significant sparsity in clinical data: for a particular combination of medical concepts, we may not find a CPT which exactly matches the given combination (i.e. $MLE(\mathcal{C})$ may be zero). For example, if the diagnoses in the CPT are $\mathcal{D}=\{$ heart attack/PRESENT, diabetes/PRESENT, obesity/ABSENT, pneumonia/POSSIBLE $\}$, we may not find any patient documented in the EMR collection who is diagnosed with all the diagnoses with the same assertions as in $\mathcal{D}$. Consequently, we would infer the likelihood of this sketch to be zero. Instead, we would prefer to consider patients whose clinical picture and therapies are *similar* to those provided in the sketch, relaxing the maximum likelihood estimation requirements. For this purpose, we considered three alternative inference techniques: (1) approximate inference based on the notion of Bethe free-energy, (2) pair-wise variational inference, and (3) inference based on interpolated smoothing.

4.1 Bethe Free-energy Approximation

We first considered state-of-the-art methods for *approximate inference*. Unlike other approaches for estimating inference, approximate inference techniques guarantee certain upper bounds on the error between their approximate probability and the true probability of the distribution. In the graphical modeling community, the most commonly used approximate inference algorithm is that of Loopy Belief Propagation [19] wherein variables and factors repeatedly exchange messages until, at convergence, the full distribution is estimated. More recently, approximate inference approaches have considered interpreting the distribution of a set of random variables as the *information energy* present in a physical system. In this setting, the distribution of all possible clinical pictures and therapies given in Equation 7 is cast as the energy J:

$$J(z)=\log\prod_{i=1}^{4}\psi_i(z)\prod_{j=1}^{6}\phi_j(z)$$

This allows us to then define the "Free Energy" of the system as follows:

$$F(z)=U(z)-H(z)=\overbrace{P(z)J(z)}^{\text{energy}}-\overbrace{P(z)\log P(z)}^{\text{entropy}} \quad (8)$$

where $U(z)$ is the energy and $H(z)$ is the entropy. As shown in [31] and [35], the minimum fixed points of the free energy equation are equivalent to fixed points of the iterative Loopy Belief Propagation algorithm. This means that minimizing the free energy in Equation 8 obtains the same solution as running iterative loopy belief propagation on Equation 7 until convergence. Moreover, the Free Energy can be approximated using the Bethe approximation which transforms our original potentially infinite message passing problem into a simple, convex, linear programming problem based on pair-wise information:

$$F_B(z,\tau)=U_B(z,\tau)-H_B(z,\tau)$$

where

$$U_B(z,\tau)=-\sum_{x\in z}\sum_{v_x\in\{0,1\}}\tau_x(v_x)\log\phi(x)$$
$$-\sum_{y\in z/\{x\}}\sum_{v_y\in\{0,1\}}\tau_{x,y}(v_x,v_y)\log\psi(x,y)$$

and

$$H_B(z,\tau) = -\sum_{x\in z}\sum_{v_x\in\{0,1\}} \tau_x(v_x)\log\tau_x(v_x) - \sum_{y\in z/\{x\}}\sum_{v_y\in\{0,1\}} \tau_{x,y}(v_x,v_y)\log\frac{\tau_{x,y}(v_x,v_y)}{\tau_x(v_x)\tau_y(v_y)}$$

Thus, we can approximate $P(z)$ from Equation 7 by minimizing F_B over τ:

$$P(z) \approx \exp\Big[-\min_{\tau} F_b(z,\tau)\Big] \quad (9)$$

where τ must satisfy the following conditions:

$$\forall x\in z, v_x\in\{0,1\} \sum_{v_y\in\{0,1\}} \tau_{x,y}(v_x,v_y) = \tau_x(v_x) \quad (10a)$$

$$\forall x\in z, y\in z/\{x\} \sum_{v_x\in\{0,1\}}\sum_{v_y\in\{0,1\}} \tau_{x,y}(v_x,v_y) = 1 \quad (10b)$$

$$\forall x\in X \sum_{v_x\in\{0,1\}} \tau_x(v_x) = 1 \quad (10c)$$

By representing the constraints in Equations 10a-10c as Lagrangian multipliers, we approximated the joint probability of any clinical picture and therapy from Equation 7 by using straightforward stochastic gradient descent[4]. In our implementation, we used the publicly available Hogwild software for parallel stochastic gradient descent [20].

4.2 Pair-wise Variational Inference

In addition to approximate inference by Bethe-free energy, we wanted to know whether simpler approximations would suffice. An obvious and much simpler strategy for relaxing the maximum likelihood estimates to better handle sparsity would be to define each factor using the association between all pairs of concepts in the factor. In this way, the four same-typed factors (ϕ_i) can be assigned to the product of all pair-wise MLE estimates in the CPTG:

$$\phi_i(z) = \prod_{x\in z}\prod_{y\in z/\{x\}} MLE(\{x,y\})$$

Likewise, the factors $\psi_1 ... \psi_6$ can be similarly defined:

$$\psi_j(z) = \prod_{u\in z}\prod_{v\in z/\{u\}} MLE(\{u,v\})$$

By using these alternative pair-wise definitions, we were able to estimate the joint distribution in Equation 7 by considering a a graph with a much simpler (i.e. pair-wise) structure. This approximation smooths the likelihood of a particular sketch by considering the likelihood of each pair of concepts in the sketch, rather than the likelihood of the entire sketch at once.

4.3 Inference Using Interpolated Smoothing

The pair-wise variational inference method defined in subsection 4.2 still suffers from sparsity problems: if the likelihood for any pair of medical concepts is zero, then the joint probability will be zero. Moreover, the pairwise approach does not discriminate between the *level of similarity* between a given CPT (e.g. z) and each CPT used to generated the CPTG. We defined the level of similarity between two CPTs as the number of concepts contained in both CPTs. Thus, the levels of similarity range from perfectly similar (all $|z|$ concepts in common) to perfectly dissimilar (0 concepts in common). In order to account for each of these levels of similarity, we interpolated the likelihood of a sketch z (or CPT), with the likelihoods of all CPTs formed by subsets of medical concepts in z. Although this would typically require enumerating all $2^{|z|}$ sub-sets of z and, thus, would be computationally intractable, we reduced the complexity to be linear in the size of the EMR collection by casting the inference problem as an information retrieval problem.

By first indexing the medical concepts present in each patient's EMRs, we were able to compute the smoothed likelihood of a particular sketch through a series of constant-time Boolean retrieval operations. Specifically, by indexing the medical concepts in the EMRs, we were able to obtain a binary vector for each medical concept in the sketch indicating which EMRs mentioned that concept. The sum of these binary vectors, which we denote as $\boldsymbol{m}$, contains, for each EMR, the number of concepts in common with the CPT of the EMR and the sketch. A single iteration over $\boldsymbol{m}$ allowed us to compute the number of EMRs with CPTs within each level of similarity denoted as $n_0 ... n_{|z|}$. The smoothed likelihood of a clinical sketch was then calculated by interpolating the number of EMRs at each similarity level (n_i):

$$P(z) \propto \alpha\cdot n_{|z|} + \sum_{i=1}^{|z|-1} (1-\alpha)^{2^{|z|-i}} n_i$$

where $\alpha \in [0,1]$ is a scaling factor such that when $\alpha = 0$ no smoothing is performed and when $\alpha = 1$, the smoothed likelihood is the sum of vector n, or the total number of EMRs whose CPT shares each level of similarity with the medical sketch.

Thus, we have considered three approaches for inferring the distribution of CPTs given by Equation 7 which is used by our system not only to rank extracted medical answers, but also to rank the documents for each answer.

5. EXPERIMENTAL RESULTS

We evaluated the role of our approach for question answering towards clinical decision support in terms of (1) the quality of answers returned for each topic as well as (2) the quality of scientific articles retrieved for each topic. In addition to the quality of answers and retrieved scientific articles, we also analyzed (3) the quality of the clinical picture and therapy graph. In these evaluations, we used the set of 30 topics (numbered 31-60) used for the 2015 TREC-CDS evaluation [25].

	$Z(t)$	$\star EZ(t,l)$
$\star$Bethe Approximation	0.125	0.694
Pair-wise Variational	0.083	0.502
Interpolated Smoothing	0.124	0.601

Table 1: The Mean Reciprocal Rank (MRR) obtained when using each of the 3 inference techniques and each medical sketch.

5.1 Medical Answer Evaluation

To determine the quality of answers produced by our system, we relied on a list of "potential answers" produced by the authors of the 2015 TREC-CDS topics and distributed by the TREC CDS organizers after the conclusion of the evaluation. These potential answers were not provided to the relevance assessors for judging document retrieval, nor were they provided to participating teams until after the evaluations were performed. Although the 2015 TREC-CDS evaluation focused only the ability to retrieve and rank scientific articles, the potential answers provided after the conclusion of the evaluation allowed us to cast the TREC-CDS task as a question-answering problem. To our knowledge, we are the first to publish any results based on the potential answers to the 30 topics used in the 2015 TREC-CDS evaluation. The potential answers indicate a single possible answer that the topic author had

[4]Although we have used stochastic gradient descent, any method for convex optimization may be used.

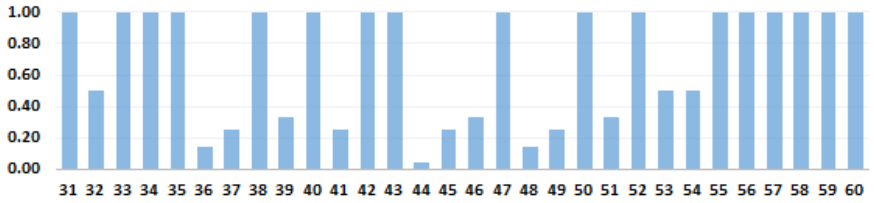

Figure 7: Reciprocal Rank for each topic evaluated in TREC-CDS 2015 based on probabilistic ranking using Interpolated Smoothing applied to the extended medical sketch $EZ(t,l)$

in mind when designing the topic; as such they do not represent the "best" answer, nor are they guaranteed to be represented in the document collection. Nevertheless, we evaluated the ranked list of answers produced by our system by using the potential answers as the gold-standard. We computed the Mean Reciprocal Rank (MRR) used in previous TREC Q/A evaluations [32], which is the average of the reciprocal (multiplicative inverse) of the rank of the first correct answer retrieved for each topic Table 1 lists the results for evaluating the inferred medical answers when considering each answer inference technique described in Section 4 for each medical sketch $z \in \{Z(t), EZ(t,l)\}$ described in Section 2.

When evaluating the answers inferred from each possible medical sketch z and answer inference method, the highest performance was obtained using the Bethe Approximation method for answer inference and relying on the extended medical sketch ($EZ(t,l)$). Note that the difference between Bethe Approximation and the Interpolated Smoothing method for probabilistic inference was not statistically significant ($p < 0.001$, $N = 30$, Wilcoxon signed-ranked test), however both methods significantly outperformed the pair-wise variational method which strongly suggests that modeling the clinical picture and therapies of patients requires more than pair-wise information about medical concepts. Overall, the answers obtained using $Z(t)$ were of significantly poorer quality than those obtained using $EZ(t,l)$. We observed that the answers produced by $Z(t)$ typically included the most common diseases, tests, or treatments indicated in the EMR collection. This confirms our belief that incorporating the medical knowledge discerned from relevant scientific articles into the sketch yields substantially higher-quality answers. The results of $EZ(t,l)$, from Table 1 show that the inferred answers from the clinical picture and therapy graph correspond reasonably-well to the possible answers generated by the TREC topic creators.

In addition to the Mean Reciprocal Rank shown in Table 1, Figure 7 shows the reciprocal rank of the gold-standard answer for each topic used in the 2015 TREC-CDS evaluation when using Interpolated Smoothing on the answers produced by $EZ(t,l)$. As shown, for the majority of topics, our top-ranked answer was equivalent to the gold-standard answer. We obtained the correct answer for nearly all of the treatment topics (topics 51-60), but for some of the diagnosis topics (31-40), and many of the test topics (41-50) we did not rank the gold answers on the first position. For example, Table 2 illustrates the ten highest-ranked answers produced by our system against the gold answer provided by the TREC topic authors, along with the held-out diagnosis for each topic previously shown in Figure 1.

As shown, for Topic 32, we obtain the correct answer at the highest position. This is because the gold answer, *cytomegalovirus* was frequently mentioned in relevant scientific articles and had a strong association to the medical case in the CPTG. Topic 44, however, was more difficult and produced the answers with the lowest MRR of any topic processed by our system. In fact, the gold answer, *flow cytometry* was ranked as the 20-th most likely answer by our system. In analyzing this behavior, we compared the answers produced by our system against a literature review of PubMed articles about *paroxysmal nocturnal hemoglobinuria*, the diagnosis for the topic. Many of the answers we proposed represent alternative

Topic 32	**EMAT:** DIAGNOSIS **Answers:** cytomegalovirus; leishmania donovani; kala-azar; mycobacterium; columbiense; salmonella; interferon-gamma; pneumonitis; lymphocytic alveolitis; pulmonary infection **Gold Answer:** cytomegalovirus
Topic 44	**EMAT:** TEST **Diagnosis:** paroxysmal nocturnal hemoglobinuria **Answers:** Hb electrophoresis; stability tests; genetic workup; renal biopsy; laboratory evaluation; ham test; sugar water tests; phosphatase; cd55; cd59; ultrasonography **Gold Answer:** flow cytometry
Topic 53	**EMAT:** TREATMENT **Diagnosis:** Dengue **Answers:** nonsteroidal anti-inflammatory drugs; fluid replacement; methylprednisolone; acetaminophen; bed rest; isotonic fluids; starch; dextran; albumin; physiotherapy; methotextrate; analgesics **Gold Answers:** supportive care, analgesics, fluid management

Table 2: Examples of answers discovered for the medical cases illustrated in Figure 1

tests recommended by Medline Plus for patients diagnosed with the disease, such as *the sugar water test*, and *Hb electrophoresis*. In contrast, the high rank of *genetic workup* highlights an area for future improvement: certain highly general tests, such as a genetic workup, are likely to have already been considered by the physician. As such, future work may benefit by narrowing the resultant answers to favor rarer diseases, tests, or treatments. For Topic 53, we obtained the correct answer as the second-highest ranked answer. This highlights the ability of our system to discover the gold answer, *fluid management*, through mentions of the synonymous concept *fluid replacement*. Unfortunately, the top-ranked answer, *nonsteroidal anti-inflammatory drugs* (NSAIDs) produced by our system was actually counter-indicated for the diagnosis of *Dengue* in most articles. That is, although the concept was mentioned with a PRESENT assertion in the article, the context indicated that the concept should not be used to treat patients with *Dengue*. This indicates that the belief values expressed by assertions may not be sufficient for answer inference in all cases.

5.2 Medical Article Retrieval Evaluation

To evaluate the quality of ranked scientific articles, we relied on the relevance judgments produced for the 2015 TREC-CDS topics by Oregon Health and Science University (OHSU). Physicians provided relevance judgments for each of the participating systems by manually reviewing the twenty top-ranked articles as well as a 20% random sample of the articles retrieved between ranks 21 and 100 for each topic. A total of 37,807 topic-article judgments were produced for the 2015 topics. These judgments indicate whether retrieved scientific articles were (1) relevant, (2) partially relevant or (3) non-relevant. In our evaluations, as in the official TREC-CDS evaluations, we did not distinguish between relevant and partially relevant documents, considering only the binary relevance of each article. This allowed us to measure the quality of articles retrieved in terms of four information retrieval metrics also used by TREC: (1) the inferred Average Precision (iAP), wherein retrieved articles were randomly sampled and the Average Precision was calculated as in [36]; (2) the inferred Normalized

Discounted Cumulative Gain (iNDCG), wherein retrieved articles were randomly sampled and the NDCG was calculated as per [37]; (3) the R-Precision, which measures the precision of the highest R-retrieved documents, where R is the total number of relevant documents for the topic; and (4) the Precision of the first ten documents retrieved (P@10) [18].

		iAP	iNDCG	R-Prec	P@10
Q/A-CDS	$Z(t)$	.006	.010	.020	.062
	$\star EZ(t,l)$	.147	.434	.344	.722
	Bethe Approximation	.140	.432	.336	.701
	Pair-wise	.128	.382	.330	.610
	$\star$Interpolated	**.147**	**.434**	**.344**	**.722**
	$\star$ BM25	.042	.204	.163	.387
	TF-IDF	.041	.197	.169	.350
	LMJM	0.040	.193	.151	.357
	LMDir	0.043	.203	.170	.360
	DFR	.039	.197	.167	.333
SotA	Task A Automatic	–	.294	–	–
	Task A Manual	–	.311	–	–
	Task B Automatic	–	.382	–	–
	Task B Manual	–	.381	–	–

Table 3: Performance results obtained for the system reported in this paper (Q/A-CDS) when using each type of medical sketch, method for answer inference, and relevance model as well as the iNDCG obtained by the State-of-the-Art (SotA) automatic and manual systems submitted to TREC.

We compared the quality of ranked scientific articles produced by our system when considering the medical sketch ($Z(t)$) or the extended medical sketch ($EZ(t,l)$), (2) each of the three answer inference methods reported in Section 4, and relying on (3) five relevance models. BM25 relied on the Okapi-BM25 [23] ($k_1=1.2$ and $b=0.75$) relevance model; TF-IDF used the standard term frequency-inverse document frequency vector retrieval relevance model; LMJM and LMDir leveraged language-model ranking functions using Jelinek-Mercer ($\lambda=0.5$) or Dirichlet ($\mu=2{,}000$) smoothing [39], respectively; and DFR considered the Divergence from Randomness framework [1] with an inverse expected document frequency model for information content, a Bernoulli-process normalization of information gain, and Zipfian term frequency normalization. We also compare our performance against the top-performing systems for the 2015 TREC-CDS evaluation for both Task A (in which no explicit diagnoses was provided) and Task B (in which an explicit diagnoses was given for each topic focusing on a medical test and treatment). It should be noted that our system did *not* incorporate the gold-standard diagnoses given in Task B (i.e. our system was designed for Task A). Moreover, our system relies on only basic query expansion (described in Section 3.1) and a standard relevance model (BM25) while the top-performing systems submitted to the TREC-CDS task relied on significantly more complex methods for query expansion and often incorporated additional information retrieval components (e.g. pseudo-relevance feedback, rank fusion) which were not considered in our architecture [22] As in the official evaluation, we distinguish between automatic systems which involved no human intervention, and manual systems in which arbitrary human intervention was allowed. Table 3 illustrates these results. Clearly, the best performance obtained by our system (denoted with a '$\star$') relies on (1) the extended medical sketch ($EZ(t,l)$), (2) the interpolated-smoothing method for answer inference, and (3) the BM25 ranking function. Note that just as with the answer evaluation, there was no statistically significant difference in the performance obtained when using the Interpolated Smoothing or the Bethe Approximation methods for answer inference. As shown, our Q/A-informed relevance approach yields significantly improved performance to top reported systems for each task [25]. In task A, we obtained a 49% increase in inferred NDCG compared to the best reported automatic system [4] and a 40% increase to the best reported manual system [4]. In task B, in which participants were given the gold-standard diagnosis for every topic (except topics 31-40 in which the purpose was to retrieve documents describing possible diagnoses), we obtained a 14% increase in inferred NDCG compared to the best reported automatic [26] and manual system [38]. This suggests that much of the increased performance obtained by our system in Task A was based on our ability to infer the correct diagnosis. Moreover, it suggests that the ability to infer multiple related medical concepts (beyond the gold-standard diagnosis) can improve the relevance of retrieved scientific articles. This suggests that the relevant articles in the TREC-CDS task considered more answers than only those in the gold-standard set. Moreover, the high performance of our approach clearly demonstrates the impact of medical Q/A for medical clinical decision support.

5.3 Medical Knowledge Evaluation

The clinical picture and therapy graph (CPTG) that we have automatically generated contained 634 thousand nodes and 13.9 billion edges with 31.2% of all nodes being diagnoses, 21.84% being signs or symptoms, 23.62% encoding medical tests, and 23.34% of nodes encoding medical treatments. The distribution of assertions that we obtained were: 13.1% were ABSENT, 0.01% were ASSOCIATED-WITH-SOMEONE-ELSE, 1.13% were CONDITIONAL, 33.31% were CONDUCTED, 17.05% were HISTORICAL, 0.72% were HYPOTHETICAL, 8.37% were ONGOING, 1.04% were ORDERED, 0.55% were POSSIBLE, 1.12% were PRESCRIBED, 22.34% were PRESENT, and 0.89% were SUGGESTED. Using the 2010 i2b2/VA shared-task annotations, medical concept detection achieved an F_1-score of 79.59%, while assertion classification obtained 92.75%. We evaluated our new assertion types using 10-fold cross validation against the 2,349 annotations we created and obtained an accuracy of 75.99%. Evaluating the quality of the edges contained in the CPTG was prevented by the fact all nodes in the CPTG are qualified by their assertions while no medical ontologies capture relations between concepts with these types of beliefs.

6. CONCLUSIONS

In this paper, a novel medical Q/A framework is presented in which answers are probabilistically inferred from an automatically derived medical knowledge graph. We experimented with three probabilistic inference methods, which enabled the identification of answers with surprisingly high MRR scores when evaluating the questions from the 2015 TREC-CDS task. Although the questions were related to complex medical cases, the results that were obtained rivaled the performance of Q/A results obtained for simpler, factoid questions. To our knowledge, the medical Q/A framework presented in this paper is the first to address the feasibility of identifying the answers to the TREC-CDS questions instead of providing a ranked list of articles from PubMed where the answers can be found. We also explored the quality of the answers obtained to the medical questions from an automatically derived medical knowledge graph in two cases: (1) when considering the medical topic by itself and (2) when considering both the medical topic and a relevant scientific article. The second case proved to be far more successful than the first one, indicating that successful medical Q/A from knowledge bases need to combine three sources of knowledge: (1) knowledge of the medical case; (2) knowledge from scientific articles (reflecting knowledge developed in medical research); and (3) knowledge from a large EMR collection (reflecting knowledge acquired during medical practice). Moreover, when the answers of a medical question are known, they inform the ranking of relevant articles from PubMed with 40%

increased inferred Average Precision to current state-of-the-art systems evaluated in the most recent TREC-CDS.

7. ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

8. REFERENCES

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

[2] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *AMIA*, page 17, 2001.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. *Dbpedia: A nucleus for a web of open data*. Springer, 2007.

[4] S. Balaneshin-kordan, A. Kotov, and R. Xisto. Wsu-ir at trec 2015 clinical decision support track: Joint weighting of explicit and latent medical query concepts from diverse sources. In *Text Retrieval Conference, TREC*, 2015.

[5] J. Bao, N. Duan, M. Zhou, and T. Zhao. Knowledge-based question answering as machine translation. *Cell*, 2(6), 2014.

[6] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267, 2004.

[7] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *International Conference on Management of Data, SIGMOD*, pages 1247–1250. ACM, 2008.

[8] W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[9] L. Dong, F. Wei, M. Zhou, and K. Xu. Question answering over freebase with multi-column convolutional neural networks. In *Association for Computational Linguistics, ACL*, volume 1, pages 260–269, 2015.

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

[11] A. X. Garg, N. K. Adhikari, H. McDonald, M. P. Rosas-Arellano, P. Devereaux, J. Beyene, J. Sam, and R. B. Haynes. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *Journal of the American Medical Association, JAMA*, 293(10):1223–1238, 2005.

[12] T. Goodwin and S. M. Harabagiu. Graphical induction of qualified medical knowledge. *International Journal of Semantic Computing, IJSC*, 7(04):377–405, 2013.

[13] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. Genia corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182, 2003.

[14] P. Kingsbury and M. Palmer. From treebank to propbank. In *LREC*. Citeseer, 2002.

[15] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[16] J. Lee, D. J. Scott, M. Villarroel, G. D. Clifford, M. Saeed, and R. G. Mark. Open-access mimic-ii database for intensive care research. In *Engineering in Medicine and Biology Society, EMBC*, pages 8315–8318. IEEE, 2011.

[17] C. E. Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[18] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[19] J. Pearl. Fusion, propagation, and structuring in belief networks. *Artificial intelligence*, 29(3):241–288, 1986.

[20] B. Recht, C. Re, S. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems, NIPS*, pages 693–701, 2011.

[21] K. Roberts and S. Harabagiu. A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association*, 18(5):568–573, 2011.

[22] K. Roberts, M. Simpson, D. Demner-Fushman, E. Voorhees, and W. Hersh. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the trec 2014 cds track. *Information Retrieval Journal*, pages 1–36, 2014.

[23] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al. Okapi at trec-3. *Text Retrieval Conference, TREC*, pages 109–109, 1995.

[24] R. H. Scheuermann, W. Ceusters, and B. Smith. Toward an ontological treatment of disease and diagnosis. *Proceedings of the 2009 AMIA Summit on Translational Bioinformatics*, 2009:116–120, 2009.

[25] M. S. Simpson, E. Voorhees, and W. Hersh. Overview of the trec 2014 clinical decision support track. In *Text Retrieval Conference, TREC*, 2014.

[26] Y. Song, Y. He, Q. Hu, and L. He. Ecnu at 2015 cds track: Two re-ranking methods in medical information retrieval. In *Proceedings of the 2015 Text Retrieval Conference*, 2015.

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

[28] M. Surdeanu and J. Turmo. Semantic role labeling using complete syntactic analysis. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 221–224. Association for Computational Linguistics, 2005.

[29] 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, JAMIA*, 18(5):552–556, 2011.

[30] H. Varmus, D. Lipman, and P. Brown. Pubmed central: An nih-operated site for electronic distribution of life sciences research reports. *Washington DC: National Institutes of Health. Retrieved December*, 24:1999, 1999.

[31] P. O. Vontobel. Counting in graph covers: A combinatorial characterization of the bethe entropy function. *Information Theory, IEEE Transactions on*, 59(9):6018–6048, 2013.

[32] E. M. Voorhees et al. The trec-8 question answering track report. In *Text Retrieval Conference, TREC*, volume 99, pages 77–82, 1999.

[33] W. A. Woods. Progress in natural language understanding: an application to lunar geology. In *Proceedings of the June 4-8, 1973, national computer conference and exposition*, pages 441–450. ACM, 1973.

[34] X. Yao and B. Van Durme. Information extraction over structured data: Question answering with freebase. In *ACL*, pages 956–966. Citeseer, 2014.

[35] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on*, 51(7):2282–2312, 2005.

[36] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 102–111. ACM, 2006.

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

[38] R. You, Y. Zhou, S. Peng, S. Zhu, and R. China. Fdumedsearch at trec 2015 clinical decision support track. In *Text Retrieval Conference, TREC.*, 2015.

[39] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Special Interest Group on Information Retrieval, SIGIR*, pages 334–342. ACM, 2001.