Knowledge Representations and Inference Techniques for Medical Question Answering

TRAVIS R. GOODWIN and SANDA M. HARABAGIU, University of Texas at Dallas

Answering medical questions related to complex medical cases, as required in modern Clinical Decision Support (CDS) systems, imposes (1) access to vast medical knowledge and (2) sophisticated inference techniques. In this article, we examine the representation and role of combining medical knowledge automatically derived from (a) clinical practice and (b) research findings for inferring answers to medical questions. Knowledge from medical practice was distilled from a vast Electronic Medical Record (EMR) system, while research knowledge was processed from biomedical articles available in PubMed Central. The knowledge automatically acquired from the EMR system took into account the clinical picture and therapy recognized from each medical record to generate a probabilistic Markov network denoted as a Clinical Picture and Therapy Graph (CPTG). Moreover, we represented the background of medical questions available from the description of each complex medical case as a medical knowledge sketch. We considered three possible representations of medical knowledge sketches that were used by four different probabilistic inference methods to pinpoint the answers from the CPTG. In addition, several answer-informed relevance models were developed to provide a ranked list of biomedical articles containing the answers. Evaluations on the TREC-CDS data show which of the medical knowledge representations and inference methods perform optimally. The experiments indicate an improvement of biomedical article ranking by 49% over state-of-the-art results.

CCS Concepts: • **Information systems → Question answering**;

Additional Key Words and Phrases: Clinical decision support, medical knowledge representation, probabilistic inference, medical information retrieval, medical question answering

ACM Reference format:
Travis R. Goodwin and Sanda M. Harabagiu. 2017. Knowledge Representations and Inference Techniques for Medical Question Answering. *ACM Trans. Intell. Syst. Technol.* 9, 2, Article 14 (October 2017), 26 pages.
https://doi.org/10.1145/3106745

1 THE PROBLEM

Physicians face in their everyday practice a variety of clinical decisions regarding the care of their patients, for example, 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 (Garg et al. 2005). 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 information needs of physicians by linking EMRs with

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.

Authors' addresses: T. R. Goodwin and S. M. Harabagiu, Department of Computer Science, University of Texas at Dallas, Richardson, TX 75083-0688, Office: ECSS-3.411; emails: {travis, sanda}@hlt.utdallas.edu.

https://doi.org/10.1145/3106745

Topic 33	Topic 42	Topic 54
EMAT: Diagnosis **Description:** A 65 yo male with no significant history of cardiovascular disease presents to the emergency room with acute onset of shortness of breath, tachypnea, and left-sided chest pain that worsens with inspiration. Of note, he underwent a right total hip replacement two weeks prior to presentation and was unable to begin physical therapy and rehabilitation for several days following the surgery due to poor pain management. Relevant physical exam findings include a respiratory rate of 35 and right calf pain. **Summary:** A 65-year-old male presents with dyspnea, tachypnea, chest pain on inspiration, and swelling and pain in the right calf. **Answer:** Pulmonary Embolism	**EMAT:** Test **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 complains of severe headache and fever. Nuchal rigidity was found on physical examination. **Diagnosis:** Bacterial Meningitis **Answer:** Spinal Tap / Cerebrospinal Fluid Analysis	**EMAT:** Treatment **Description:** A 31 yo male with no significant past medical history presents with productive cough and chest pain. He reports developing cold symptoms one week ago that were improving until two days ago, when he developed a new fever, chills, and worsening cough. He has right-sided chest pain that is aggravated by coughing. His wife also had cold symptoms a week ago but is now feeling well. Vitals signs include temperature 103.4, pulse 105, blood pressure 120/80, and respiratory rate 15. Lung exam reveals expiratory wheezing, decreased breath sounds, and egophany in the left lower lung field. **Summary:** A 31 year old male presents with productive cough, chest pain, fever and chills. On exam he has audible wheezing with decreased breath sounds and dullness to percussion. **Diagnosis:** Community Acquired Pneumonia (CAP) **Answer:** Antibiotics

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

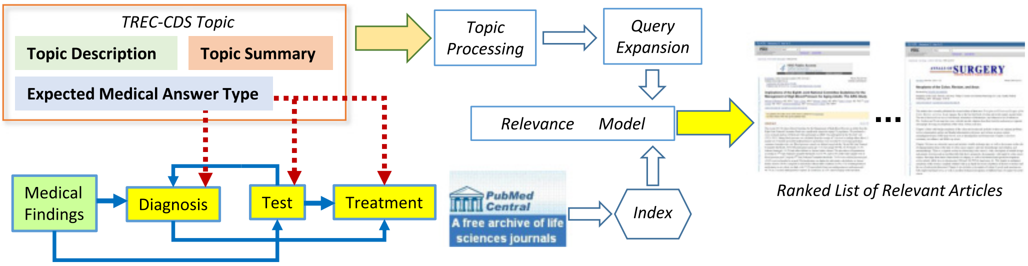

Fig. 2. Architecture of a typical medical question answering system for clinical decision support.

information relevant for patient care, retrieved from the bio-medical literature. Recently, the special track on Clinical Decision Support in the Text REtrieval Conference (TREC-CDS) (Simpson et al. 2014) has addressed the challenge of retrieving bio-medical articles relevant 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, but instead it used an idealized representation of medical records in the form of 30 short *medical case reports*, each describing a challenging medical case. The medical case reports were presented in two formats: (a) a narrative describing the fragments from the patient’s EMRs that were pertinent to the case detailed description or (b) a summary of the case. Medical case reports (in both formats) along with one of the generic questions were considered *topics* in TREC-CDS. Thus, systems developed for the TREC-CDS challenge were provided with a list of topics and 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 (Varmus et al. 1999) 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.

Most systems that participated in TREC-CDS used architectures similar to the one illustrated in Figure 2. The topics were processed to discover terms or concepts that were used to generate a query. Queries were expanded to enhance the quality of retrieval enabled by a relevance model operating on the index of the PubMed Central collection. Roberts et al. (2016) details the variety of topic processing, query expansions, and relevance modes used by the various systems. Notably, Roberts et al. (2016) also discusses the relations between the three types of expected medical answers (EMATs), namely the diagnosis, the tests, and the treatments as well as the medical findings pertaining to the difficult medical case addressed by the CDS topics. Figure 2 also illustrates these relationships which influence the clinical decision making process, as stated in Roberts et al. (2016). Medical findings mentioned in the medical case description and/or summary require the inference of a diagnosis (differential or confirmed) and require tests to be ordered, either preliminary or confirmatory. The tests may confirm the diagnosis, thus an additional relation exists between the tests and the diagnosis. Treatments are dependent on both the diagnosis and the results of the tests. As Figure 2 illustrates, the EMAT of a CDS topic cannot be considered in isolation, given the dependencies between the three EMATs considered in TREC-CDS and the medical findings pertaining to the medical case. Moreover, we believe that these dependencies should be further explored, an insight that was not considered by any of the systems participating in the TREC-CDS.

In the research presented in this article, we extend the work reported in Goodwin and Harabagiu (2016) by focusing on the medical knowledge representations that can be successfully considered in medical question answering (Q/A). We contemplated knowledge representations in which the EMATs of questions evaluated in TREC-CDS are considered along with additional medical concepts, as well the connections shared by them. However, we constrained these representations to take into account connections between medical concepts that are observed in medical practice and thus are infer-able from a vast EMR collection. Moreover, we considered a probabilistic representation of the medical knowledge, in which answers to the medical questions like those evaluated in TREC-CDS can be inferred instead of only being searched. Thus, we believed that by focusing on medical knowledge representations and experimenting with inference methods operating on them, we could not only find the optimal knowledge representations and inference methods for medical Q/A, but we could also further improve the medical article retrieval results reported in Goodwin and Harabagiu (2016), as it would allow us to produce new answer-informed relevance models. Our belief that improved answer inference must lead to enhance article retrieval originated from observations of results of the 2015 TREC-CDS.

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). In this way, some of the dependencies related to clinical decision were exposed. The results for this new task, as reported in Roberts et al. (2015) were superior to the results for the same topics when no diagnoses were provided. This observation led us to believe that medical knowledge related to the EMAT can be considered as a partial answer to the medical question of the CDS topic. The results from the evaluation of the new TREC-CDS task indicated that knowledge of a partial answer leads to significantly improved retrieval of relevant bio-medical literature. Thus, we asked ourselves if (1) we could automatically assemble medical knowledge that could provide partial answers for any CDS topic; and (2) if we could in fact identify the answers to medical questions from the CDS topics with acceptable accuracy, if such medical knowledge would be available. More importantly, we wondered if we should first try to find the answer and then rank the relevant scientific articles for any given medical question.

As first reported in Goodwin and Harabagiu (2016), 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 1960s and '70s, domain-specific knowledge bases were used to support Q/A, e.g. the Lunar Q/A system (Woods 1973). With the recent growth of KBs such as DBPedia (Auer et al. 2007) and Freebase (Bollacker et al. 2008), new promising methods for Q/A from KBs have emerged (Dong et al. 2015; Yao and Van Durme 2014; Bao et al. 2014). These methods map questions into sophisticated meaning-representations that are used to retrieve the answers from the KB. However, we believe that when probabilistic representations of the medical knowledge are available, medical Q/A from KB can achieve significant accuracy when the inference methods take advantage of the various forms of medical knowledge. It is equally important to capture the background of medical questions in medical knowledge sketches. We considered three different forms of medical knowledge sketches that combine in different ways the knowledge processed from the description of the medical case with knowledge processed from clinical practice and medical research. Therefore, in this article, we expand the work originally reported in Goodwin and Harabagiu (2016) by exploring three different possible medical knowledge sketches and four different inference methods to discover which combination of knowledge representation and inference approach produces optimal results of medical Q/A on TREC-CDS data.

The ability to cast the problem of answering medical questions for CDS topics as a Q/A from KB problem depends on the availability of a large medical knowledge base in which the *clinical picture* (comprising the medical findings and the diagnoses) and *therapy* (comprising the tests and treatments) of a vast population of patients is captured. Moreover, in this medical knowledge base, the dependencies between diagnoses, medical findings, tests, and treatments would be not only available but also captured at the level of each patient and medical case, providing the knowledge granularity required by the medical questions evaluated in TREC-CDS (as illustrated in Figure 2). To our knowledge, no such knowledge base is readily available. Widely used medical ontologies such as the Unified Medical Language System (UMLS) (Bodenreider 2004) and MeSH (Lipscomb 2000) encode a large number of medical concepts, but these ontologies do not relate medical concepts to any specific medical case. We believe that a medical knowledge base that could inform the medical questions asked in TREC-CDS should capture the knowledge documented in a vast collection of medical records. For this purpose, we automatically generated a very large medical knowledge graph from a publicly available collection of electronic medical records (EMRs). Because, as reported in Roberts et al. (2016), the medical case descriptions from the TREC-CDS topics were generated by consulting the EMRs from MIMIC-II (Lee et al. 2011), 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 knowledge graph designed to encode knowledge acquired from medical practice.

We organized the medical knowledge acquired from the EMR collection into a clinical picture and therapy graph (CPTG), which informed, along with the medical knowledge sketches resulting from topic processing, answer inference—the central component of the new medical Q/A system illustrated in Figure 3. The answers, ranked by their likelihood, enable novel answer-informed relevance models to produce a ranked list of relevant biomedical articles, based on the index of PubMed Central.

In the architecture illustrated in Figure 3, the CPTG is automatically generated by (1) processing the language from the narratives of medical records to identify medical concepts (and their assertions); and (2) inferring probabilistically the connections between medical concepts. TREC-CDS topics can be processed to discern medical concepts and their assertions in the same format as the one used in the CPTG. By identifying (a) several forms of medical concepts including signs and symptoms, diagnoses, as well as tests and treatments; and (b) the way in which they are asserted (e.g. PRESENT, ABSENT, POSSIBLE, etc.), we considered a richer semantic representation of medical knowledge that the one provided by the three EMATs on the medical finding discussed in Roberts et al. (2015). Furthermore, in this article, we consider for the CPTG more complex representations of connections between medical concepts than those discussed in Roberts et al. (2016). Answers

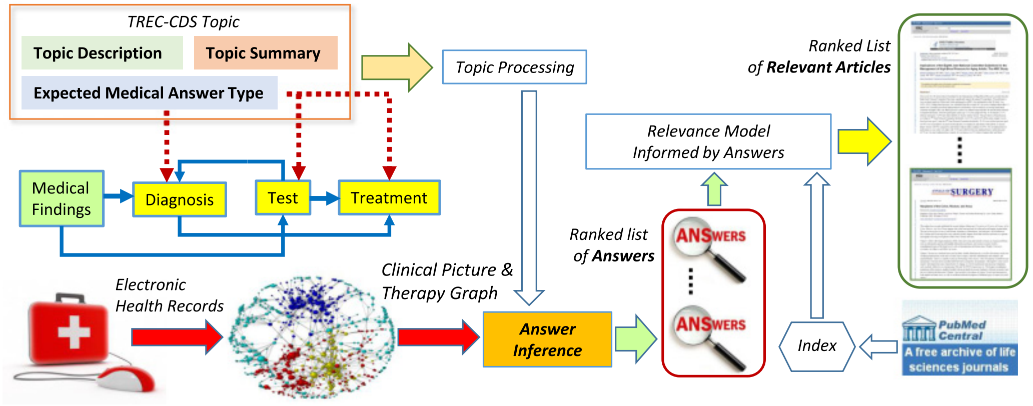

Fig. 3. Architecture of our medical question answering system for clinical decision support.

inferred from the CPTG were evaluated against correct answers made available to all participants in the 2015 TREC-CDS challenge. In addition, evaluations of relevant biomedical articles could be performed for the new form of medical Q/A, first reported in Goodwin and Harabagiu (2016) and illustrated in Figure 3. The results confirmed our intuition that it is not necessary to first discover relevant biomedical articles from which the answers can be extracted (as was the case with previous textual Q/A systems), but we could infer the answers directly from the CPTG (i.e., the medical KB) and then discover the relevant articles that contained the already known answers to provide additional information and context for the answers. In the extended research reported in this article, we explore the ideal knowledge representations of medical knowledge that can be used as well as the ideal inference methods than can be considered in an medical Q/A system similar to the architecture illustrated in Figure 3. Furthermore, we experimented with another lesson learned for textual Q/A, namely to search for an answer in texts by preferring one paragraph at a time instead of the entire document. Our experiments show that this lesson is not applicable to medical Q/A, as knowledge distilled only from one paragraph leads to inference of less accurate answers. This may be explained by the fact that scientific articles organize knowledge about a complex medical case across multiple paragraphs, and thus the knowledge distilled from only one paragraph is insufficient.

In this article, we present the knowledge representations considered for medical Q/A used in clinical decision support as well as the details of the probabilistic inference methods that were used, providing the following main contributions:

1. A probabilistic representation of the medical knowledge processed from a vast EMR collection known as a Clinical Picture and Therapy Graph (CPTG). The CPTG is represented as a Markov network which implements: (i) nodes for diagnoses, tests, treatments, symptoms and signs; and (ii) factors for the connections between them;
2. Usage of assigned and latent random variables in the CPTG to account for medical concepts explicitly expressed or inferable, respectively;
3. Using the likelihood of the automatically discovered answers to produce several novel answer-informed rankings of the relevant scientific articles;
4. Locating the answers both at document- and paragraph-level and showcasing the impact that paragraph indexing produces on the quality of article relevance; and
5. Designing, implementing, and evaluating a system for answering medical questions which can consider (a) the three medical knowledge sketches as well as (b) four different probabilistic inference methods for discovering answers from the CPTG.

The remainder of the article is organized as follows. Section 2 details the new architecture for answering medical questions. Section 3 describes the CPTG used for capturing the necessary medical knowledge and details the probabilistic inference methods that were used for discovering answers. Section 4 details the methods used for automatically generating the CPTG, while Section 5 presents and discusses the experimental results. Section 6 summarizes the conclusions.

2 AN ARCHITECTURE FOR INFERRING MEDICAL ANSWERS

The design of a Q/A architecture that operates on TREC-CDS topics and provides both answers and relevant biomedical answers from PubMed needs to take into account (a) the medical knowledge base that informs the answer inference as well as (b) the multiple ways in which, once the answer is discovered, new relevance models can identify and rank the relevant PubMed articles and compare them to the relevance results obtained when the answers are ignored.

2.1 Inferring Answers for the TREC-CDS Topics by Using Medical Knowledge Sketches

The cornerstone of our medical Q/A method used 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 concept (and its belief value). We automatically identified four types of medical concepts: signs/symptoms, tests, diagnoses, and treatments (as detailed in Section 4.1). 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 4.2. In Section 4, we describe the methods used to identify automatically medical concepts and their associated assertions, reflecting the belief values of the physician that authored the clinical narrative. Section 3.1 details the estimations used in the CPTG, while Section 3.2 describes the inference methods used for discovering the answers to TREC-CDS topics.

However, answers to the TREC-CDS medical questions need also to (1) account for the medical knowledge expressed in the description and/or summary of TREC-CDS topics; and (2) be mentioned in relevant biomedical articles from PubMed Central. For the first desideratum, for any of the TREC-CDS topics, t, a medical knowledge sketch $Z_1(t)$ was discerned. $Z_1(t)$ accounts for the clinical picture and therapy of the complex medical case referred by the topic t and it consists of medical concepts and their inferred assertions. The second desideratum constrained us to discover only answers to the question from t that can also be observed in biomedical articles from PubMed Central. Thus, to find such answers, for any biomedical document l (from PubMed Central) that was relevant to the question from topic t, we considered the medical knowledge sketch $Z_2(t, l)$, which combined $Z_1(t)$ with medical concepts (qualified by their assertions) recognized, l. We believe that $Z_2(t, l)$ accounts for a more complete view of a possible clinical picture and therapy of a medical case than the one discerned only from the topic. This belief was strengthened by the observation that the joint distribution estimated from the CPTG favors more common medical concepts, whereas the topics evaluated in the TREC-CDS correspond to complex medical cases, rather than common cases.

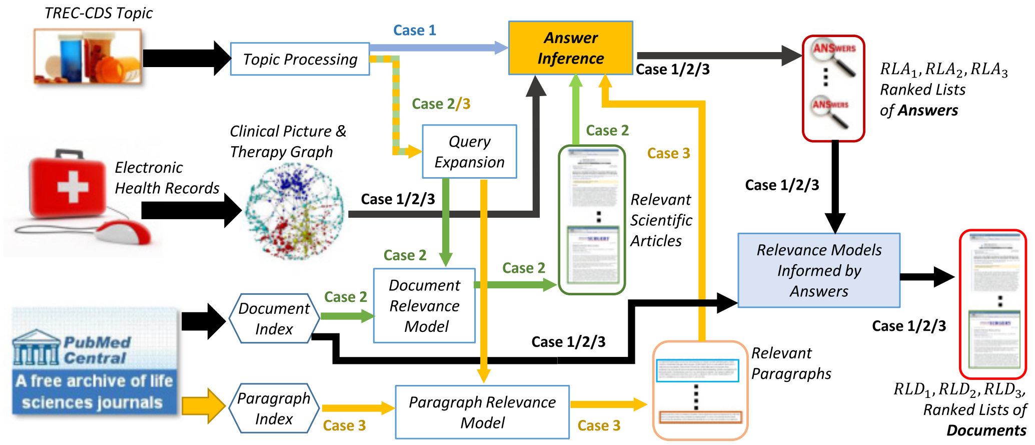
TREC-CDS Topic
Topic Processing
Case 1
Answer Inference
Case 1/2/3
Case 2/3
ANSWERS
RLA$_1$, RLA$_2$, RLA$_3$ Ranked Lists of Answers
Clinical Picture & Therapy Graph
Query Expansion
Case 2
Case 3
Electronic Health Records
Case 1/2/3
Relevant Scientific Articles
Case 1/2/3
Relevance Models Informed by Answers
Case 2
Case 2
Document Index
Case 2
Document Relevance Model
Case 1/2/3
Case 1/2/3
PubMed Central
A free archive of life sciences journals
Relevant Paragraphs
Paragraph Index
Case 3
Paragraph Relevance Model
Case 3
RLD$_1$, RLD$_2$, RLD$_3$, Ranked Lists of Documents

Fig. 4. An architecture that implements three different cases for answering medical questions for clinical decision support.

We also wondered if an alternative medical knowledge sketch, which adds to $Z_1(t)$ only qualified medical concepts recognized within the same paragraph of biomedical documents relevant to the topic would not enable the inference of better answers. Textual Q/A is known to produce superior results when answers are extracted from paragraphs rather than documents (Harabagiu and Maiorano 1999). Thus, we also build a medical knowledge sketch denoted as $Z_3(t, s)$ that adds to $Z_1(t)$ only asserted medical concepts from a paragraph s, belonging to a biomedical document from PubMed Central known to be relevant to the question of the topic t. In this way, we were able to consider three forms of knowledge sketches, namely $Z_1(t)$, $Z_2(t, l)$, or $Z_3(t, s)$, for identifying the answer to a TREC-CDS topic. If $z \in \{Z_1(t), Z_2(t, l), Z_3(t, s)\}$ is any of the medical knowledge sketches, then we could discover the most likely answer $\hat{a}$ to the medical question associated with t by discovering the medical concept encoded in the CPTG, which, when combined with the sketch, produces the most likely clinical picture and therapy. Formally,

$$\hat{a} = \underset{a \in A}{\operatorname{argmax}} \, P\left(a|z\right) = \frac{P\left(\{a\} \cup z\right)}{P\left(z\right)}, \tag{1}$$

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

2.2 Architecture of Medical Q/A System used in Clinical Decision Support

The architecture of the medical QA system that can be used in Clinical Decision Support (CDS) is illustrated in Figure 4. We envisioned three cases or scenarios for using this architecture:

□ *Case 1* infers answers from the CPTG only and uses the answers to retrieve PubMed Central documents relevant to the question implied by the topic's EMAT (e.g., *what is the most likely diagnosis/treatment/test?*);

□ *Case 2* combines the advantages of a vast medical knowledge base (provided by the CPTG) with the document relevance model to infer answers, which are later used by an answer-informed relevance model (different from the one used in case 1) to identify PubMed Central documents relevant to the question implied by the topic's EMAT; and

□ *Case 3* combines the same medical knowledge base used in cases 1 and 2 with a paragraph relevance model to infer the answers, which are later used by an answer-informed relevance model (different from the ones used in cases 1 or 2) to identify PubMed Central documents relevant to the question implied by the topic's EMAT.

Details of each of these cases are provided below.

Case 1. In case 1, as illustrated in Figure 4, the topic is processed with methods detailed in Section 4 to discern the medical concepts mentioned in the topic's description or summary and to discover their assertions. Topic processing generates the medical knowledge sketch $Z_1(t)$, which is used to produce the ranked list of answers RLA_1 according to Equation (1) based on inference enabled by the CPTG. Additionally, we designed an answer-informed relevance model to also identify the ranked list of documents RLD_1, relevant to the medical question from each topic. When a document l_i from PubMed Central (retrieved from the document index), contains an answer $Y_i \in RLA_1$, we defined the answer-informed relevance to the topic t by

$$\mathrm{Rel}(l_i) = P(Y_i | Z_1(t)) \propto P(Y_i \cup Z_1(t)). \tag{2}$$

Equation (2) represents an answer-informed ranking of each scientific article l_i that contains answers Y_i from RLA_1 based on the likelihood of the answers in the article, given the medical knowledge sketch derived for the topic.

Case 2. In the case 2 illustrated in Figure 4, topic processing is also used to produce a query (as described below), which can be further expanded, as it was typically done in the systems participating in the TREC-CDS challenge (and illustrated in Figure 2). 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) (Bodenreider 2004) and MeSH (Lipscomb 2000) were commonly used as ontological resources for medical concepts. When generating the query, we opted to use medical concepts rather than words and formed a disjunctive Boolean query by considering each medical concept as a key phrase. When the query was expanded, additional medical concepts from UMLS that share the same concept unique identifier (CUI) as any medical concept detected from the topic were added. These concepts represent synonyms of the topic concepts; thus, the same assertion was extended to them as well. It should be noted that some medical concepts (e.g., "heart failure") or their synonyms (e.g., "cardiac insufficiency") are phrases rather than single words. Consequently, the resulting expanded query consists of a list of key-phrases representing medical concepts. For example, the initial Boolean query produced for Topic 33 (from Figure 1) would include "cardiovascular disease" OR "shortness of breath" OR "tachypenea," and so on, while the expanded Boolean query would include "cardiovascular disease" OR "cardiovascular disorder" OR "CVD," and so on. The document relevance model illustrated in Figure 4 makes use of the expanded query, the index of the PubMed Central articles, and a document relevance model to retrieve a list L of 1,000 relevant documents. In the index, we used a snapshot of PubMed Central articles from January 21, 2014 containing a total of 733,138 articles that were provided by the TREC-CDS organizers. In TREC-CDS, systems implemented a variety of relevance models, as reported in Roberts et al. (2015), to generate the ranked list of relevant articles. We also experimented with several document relevance models (discussed in Section 5) to retrieve the first 1,000 most relevant documents, denoted as L. We used L to produce the medical knowledge sketch $Z_2(t, l)$, for each topic t and relevant document $l \in L$.

A close inspection of the contents of the medical knowledge sketches $Z_2(t, l)$ indicated the inclusion of many medical concepts obtained from scientific articles presented no relevance to the topic. This reflects the fact that many of the scientific articles in PubMed Central discuss unexpected or unusual medical cases—often in non-human subjects. This created a serious problem in the usage of the medical knowledge sketch $Z_2(t,l)$ to infer answers from the CPTG to the question from a topic t. Specifically, because the likelihood estimate of an answer enabled by the CPTG is based on the observed clinical pictures and therapies of patients documented in the MIMIC clinical database, non-relevant scientific articles that contained common diagnoses, treatments, tests, signs, or symptoms had a disproportionately large impact on the ranking of answers. To address this problem, we refined the ranking of answers provided by Equation (1) to incorporate the relevance of the scientific article $l \in L$ used for generating the medical knowledge sketch $Z_2(t,l)$. Thus, in case 2, we produced the answer ranking by using a novel probabilistic metric, namely the Reciprocal-Rank Article Score (RRAS). RRAS considers for each article $l_r \in L$: (1) the conditional probability of the answer given the medical knowledge sketch $Z_2(t,l_r)$; as well as (2) the relevance rank, r of the article l_r in L. Formally, the new ranking of answers to a question associated with topic t generated by the RRAS metric is defined as

$$\mathrm{RRAS}\,(a) = \sum_{r=1}^{1{,}000} \frac{1}{r} \cdot P\,(a|Z_2\,(t,l_r)) = \frac{P\,(\{a\} \cup Z_2\,(t,l_r))}{P\,(Z_2\,(t,l_r))}. \tag{3}$$

The list of ranked answers, denoted as RLA_2, was ranked by using the RRAS metric. The document index was used to retrieve the set of PubMed Central documents that contain answers from RLA_2. These documents were retrieved by using a Boolean query that used a disjunction of all medical concepts from $Z_2(t,l)$. However, this document set needed to be ranked to produce the ranked list of documents RLD_2, relevant to the medical question from each topic in case 2. We defined a new, answer-informed relevance model, which ranked the documents from RLD_2. Specifically, when a document l_i from RLD_2 contains answers $Y_i \in RLA_2$, the answer-informed relevance of l_i to the topic t is provided by

$$\mathrm{Rel}\,(l_i) = P\,(Y_i|Z_2\,(t,l_i)) = \frac{P\,(Z_2\,(t,l_i))}{P\,(Z_2\,(t,l_i) - Y_i)}. \tag{4}$$

In this way, the relevance of an article l_i responding to the question of a topic is computed by comparing the likelihood of the medical knowledge sketch $Z_2(t,l_i)$, which includes the answers found in the article l_i against the likelihood of a version of the medical knowledge sketch $Z_2(t,l_i)$, which does not contain the answers found in the article.

Case 3. Finally, in case 3, answers are inferred from the CPTG using $Z_3(t,s)$. Thus, a new set of answers are inferred (details of inference methods are provided in Section 3) and ranked in a new answer list denoted as RLA_3. In RLA_3, ranking is provided by a Reciprocal-Rank Paragraph Score (RRPS) defined as

$$\mathrm{RRPS}\,(a) = \sum_{r=1}^{1{,}000} \frac{1}{r} \times \max_{s_k \in l_r} P\,(a|Z_3\,(t,s_k))\,, \tag{5}$$

where the article l_r has the rank r in L and paragraph s_k indicates the kth paragraph in article l_r. In the definition of the ranking metric RRPS used in RLA_3, we took into account (a) the rank r of the article from L that contained the answer; as well as (b) the most likely paragraph from the same article that contained the answer. Furthermore, the ranking generated by RPPS on the list of answers RLA_3 was used by yet another answer-informed relevance model to produce RLD_3, the ranked list of documents from PubMed Central relevant to the question from a TREC-CDS topic. The ranking in this list of articles is generated by

$$\mathrm{Rel}\,(l_i) = \max_{s_k \in l_i} P\,(Y_i|Z_3\,(t,s_k))\,, \tag{6}$$

Table 1. Overview of the Different Medical Knowledge Sketches, Document Relevance Models, and Answer Ranking Metrics Used in Each Case Illustrated in Figure 4

Case	Sketch	Document Relevance Model	Answer Ranking Metric
Case 1	$Z_1(t)$	$P(Y_i \| Z_1(t))$ (Equation (2))	$P(a\|z)$ (Equation (1))
Case 2	$Z_2(t,d)$	$P(Y_i \| Z_2(t,l_i)$ (Equation (3))	RRAS (a) (Equation (4))
Case 3	$Z_3(t,s)$	$P(Y_i \| Z_3(t,s_k)$ (Equation (6))	RRPS (a) (Equation (5))

where Y_i represents all answers from RLA_3 found in an article ranked on position i of L. Because not all answers from Y_i may be found in each paragraph of the article l_i, the ranking favors the articles that (a) contain most of the answers in a single paragraph; and (b) also contain most of the concepts from the topic in the same paragraph.

In summary, the architecture illustrated in Figure 4 enabled us to use medical Q/A in CDS system in three different cases, summarized in Table 1. Each case generated a different list of ranked answers (RLA_1, RLA_2, and RLA_3, depending on the usage of medical knowledge sketches $Z_1(t)$, $Z_2(t,l)$, or $Z_3(t,s)$) as well as a different list of relevant articles for each topic, namely RLD_1, RLD_2, and RLD_3, produced by three different answer-informed relevance models.

3 ANSWER INFERENCE

In this section, we first describe the ontological principles used in representing the medical knowledge as a Markov network, a special form of probabilistic graphical models. We detail how medical concepts and their inferred assertions are represented, as well as how relations between the various medical concepts are considered. We present the way in which any combination of medical concepts (including the medical knowledge sketches used for answering medical questions in the architecture presented in Figure 4) can be represented such that answer inference can be achieved through Equations (1)–(6). In the second part of the section, we describe four distinct inference methods that were used for inferring answers to the TREC-CDS questions from the CPTG.

3.1 Medical Knowledge Representation in the Clinical Picture and Therapy Graph

The ontological framework introduced in Scheuermann et al. (2009) considers (1) the *clinical picture* of a patient, consisting of the medical problems, signs/symptoms, and medical tests that might influence the diagnosis of the patient; and (2) the *therapy* of a patient, consisting of the set of all treatments, cures, and preventions included within the management plan for the patient. Medical language processing methods detailed in Section 4 enable us to discern from a vast EMR collection the medical concepts that represent the clinical picture and therapy (CPT) mentioned in the narrative of each medical record. Moreover, the methods described in Section 4 allow us to associate with each medical concept an assertion value, indicating whether the medical concept is PRESENT, ABSENT, and so on (the full list of assertion values and their definitions is provided in Section 4.2). Thus, a CPT lists the set of medical concepts from the same medical record, with each concept having its own assertion. It is important to note that the CPT varies significantly between patients with the same disease (e.g., in one patient, a symptom may be PRESENT, whereas in another ABSENT) and often varies across different points in time for the same patient during the course of their care (e.g., a patient that had fever PRESENT at some point, after a treatment with antibiotics, the fever resolves and is asserted as ABSENT).

The medical knowledge acquired automatically from the EMR collection was represented in the CPTG such that each node corresponds to a medical concept observed in any CPT derived from the EMR collection. We decided to represent the CPTG as a factorized Markov network (Koller and

Friedman 2009) in which the nodes are partitioned into: (1) $\mathbb{D}$ representing all the diagnoses; (2) $\mathbb{S}$ representing all the signs/symptoms; (3) $\mathbb{E}$ representing all the tests; and (4) $\mathbb{R}$ representing all the medical treatments. It is important to note that factorized Markov networks encode knowledge by using (i) statistical random variables and (ii) mathematical factors (or functions) measuring the strength of the relationships between statistical random variables in the model. Hence, in the CPTG, each medical concept (i.e., each node) is a binary random variable that is assigned the value of 1 when the medical concept was asserted to be PRESENT, CONDUCTED, ORDERED, or PRESCRIBED, a value of 0 if the medical concept was asserted as ABSENT, and was left as a *latent* or unassigned variable, otherwise.

Given a CPT derived from a medical record, it is important to note that (a) each medical concept from the CPT has a random variable assigned to it; and (b) all the nodes from the CPTG not recognized in the CPT are associated with latent variables, whose values can be later inferred during answer inference. It is important to note that by relying on both assigned and latent random variables, we enable the representation of any possible combination of medical concepts (and their assertions), regardless of whether they were mentioned in the same electronic medical record or not. Possible combinations of medical concepts include the three medical knowledge sketches that we have considered in the architecture for answering medical questions that can be used in clinical decision support, defined in Section 2.1, namely $Z_1(t)$, $Z_2(t, l)$, and $Z_3(t, s)$. All of these medical knowledge sketches consist of medical concepts and their assertions. The probabilistic representation of any of these medical knowledge sketches consists of (1) random variables assigned to the medical concepts identified in the sketch; and (2) latent variables corresponding to all the other nodes from the CPTG.

Formally, a combination of medical concepts (and their assertions) is denoted by $\mathcal{C} = \mathcal{D} \cup \mathcal{S} \cup \mathcal{E} \cup \mathcal{R}$, where $\mathcal{D} \subseteq \mathbb{D}$ indicates the random variables corresponding to the diagnoses in $\mathcal{C}$, $\mathcal{S} \subseteq \mathbb{S}$ indicates the random variables corresponding to signs/symptoms in $\mathcal{C}$, $\mathcal{E} \subseteq \mathbb{E}$ indicates the random variables corresponding to tests in $\mathcal{C}$, and $\mathcal{R} \subseteq \mathbb{R}$ indicates the random variables corresponding to treatments in $\mathcal{C}$. The estimation of the probability of any combination of medical concepts (and their assertions) $\mathcal{C}$ is made possible once the factors of the CPTG (represented as a Markov network) are defined. We defined ten factors. The first four factors are: $\phi_1(\mathcal{D})$, the likelihood of a CPT (derived from a medical record) containing the diagnoses from $\mathcal{D}$; $\phi_2(\mathcal{S})$, the likelihood of a CPT (derived from a medical record) containing the signs/symptoms from $\mathcal{S}$; $\phi_3(\mathcal{E})$, the likelihood of a CPT (derived from a medical record) containing the tests from $\mathcal{E}$; and $\phi_4(\mathcal{R})$, the likelihood of a CPT (derived from a medical record) containing the treatments from $\mathcal{R}$. In addition, the CPTG encodes relations between medical concepts of different types. Six additional factors enable the probabilistic representation of these relations: (1) $\psi_1(\mathbb{D}, \mathbb{S})$, the strength of the correlation between all the diagnoses in $\mathbb{D}$ and all the signs/symptoms in $\mathbb{S}$; (2) $\psi_2(\mathbb{S}, \mathbb{E})$, the strength of the correlation between all the signs/symptoms in $\mathbb{S}$ and all the tests in $\mathbb{E}$; (3) $\psi_3(\mathbb{D}, \mathbb{E})$, the strength of the correlation between all the diagnoses in $\mathbb{D}$ and all the tests in $\mathbb{E}$; (4) $\psi_4(\mathbb{D}, \mathbb{R})$, the strength of the correlation between all the diagnoses in $\mathbb{D}$ and all the treatments in $\mathbb{R}$; (5) $\psi_5(\mathbb{E}, \mathbb{R})$, the strength of the correlation between all the tests in $\mathbb{E}$ and all the treatments in $\mathbb{R}$; and (6) $\psi_6(\mathbb{S}, \mathbb{R})$, the strength of the correlation between all the signs/symptoms in $\mathbb{S}$ and all the treatments in $\mathbb{R}$. It is to be noted that, unlike knowledge graphs that typically encode binary relations, the factors used in the CPTG correspond to *hyper-edges* representing n-ary relations between many nodes in the graph. Figure 5 illustrates the representation of the CPTG as a factorized Markov network in which each partition of random variables $\mathbb{D}$, $\mathbb{S}$, $\mathbb{E}$, and $\mathbb{R}$ is shown. Figure 5 also illustrates the representation of a combination of medical concepts (and their assertions), $\mathcal{C} = \mathcal{D} \cup \mathcal{S} \cup \mathcal{E} \cup \mathcal{R}$. As such, the random variables with assigned values in $\mathcal{D}$, $\mathcal{S}$, $\mathcal{E}$, or $\mathcal{R}$ are represented as filled circles, whereas the latent variables corresponding to nodes from the CPTG that are not present in $\mathcal{C}$ are represented

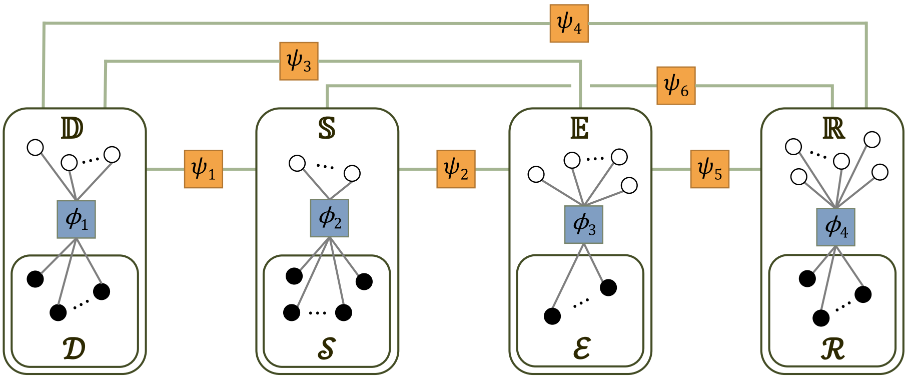

Fig. 5. Factorized Markov network representation of the Clinical Picture and Therapy Graph (CPTG) and the representation of any combination of medical concepts (and their assertions) $C = \mathcal{D} \cup \mathcal{S} \cup \mathcal{E} \cup \mathcal{R}$ in the CPTG.

as empty circles. Figure 5 also illustrates the factor ϕ_1 in $\mathbb{D}$ as an n-ary relation involving all the assigned and latent variables in $\mathbb{D}$. Similarly, factors ϕ_2, ϕ_3, and ϕ_4 encode n-ary relations involving all the random variables in $\mathbb{S}$, $\mathbb{E}$, and $\mathbb{R}$, respectively. Moreover, the factors ψ_1, ψ_2, ψ_3, ψ_4, ψ_5, and ψ_6, which encode the n-ary relations between medical concepts of different types, are also shown.

The factorized Markov network representation of the CPTG illustrated in Figure 5 enables us to compute the probability of any combination of medical concepts (and their assertions) C using:

$$P(C) = P(\mathcal{D}, \mathcal{S}, \mathcal{E}, \mathcal{R}) \propto \phi_1(\mathcal{D}) \times \phi_2(\mathcal{S}) \times \phi_3(\mathcal{E}) \times \phi_4(\mathcal{R}) \times \psi_1(\mathcal{D}, \mathcal{S}) \times \psi_2(\mathcal{S}, \mathcal{E}) \times \psi_3(\mathcal{D}, \mathcal{E}) \\ \times \psi_4(\mathcal{D}, \mathcal{R}) \times \psi_5(\mathcal{E}, \mathcal{R}) \times \psi_6(\mathcal{S}, \mathcal{R}). \tag{7}$$

The probability distribution provided in Equation (7) is determined by the product of the ten factors used in the factorized Markov network representation of the CPTG. Unfortunately, evaluating any of these factors directly can be intractably expensive. For example, pre-computing $\psi_1(\mathcal{D}, \mathcal{S})$ requires storing $2^{|\mathcal{D}| \times |\mathcal{S}|}$ probabilities (or counts). Beyond computational complexity, an additional problem arises from the inherent sparsity of clinical data: for a given combination of medical concepts, it is very unlikely that the given combination may be reflected in the CPT of any patient in the EMR collection, thus, the probability assigned to that combination would be zero. For example, if the diagnoses in C are $\mathcal{D} = \{$[heart attack/PRESENT], [diabetes/PRESENT], [obesity/ABSENT], [pneumonia/POSSIBLE]$\}$, we may not find any patient documented in the EMR collection who has all the diagnoses with the same assertions as in $\mathcal{D}$. Consequently, we would infer the likelihood of C as zero. If C were a medical knowledge sketch, then we would not find any answers to the medical question of a TREC-CDS topic. To address this problem, we decided to consider EMRs with narratives describing clinical pictures and therapies that are *similar* to C (e.g., any of the medical knowledge sketches). To do so, we relaxed the maximum likelihood estimation requirements. This allowed us to infer the ten factors used to compute $P(C)$ using four methods: (1) exact inference, (2) pair-wise smoothing, (3) interpolated smoothing, and (4) applying the Bethe free-energy approximation. When inferring the factors with the methods described in Section 3.2, we are able to use Equation (7) to compute the probability of any given combination of medical concepts as required for ranking both answers and scientific articles (Equations (1)–(6)).

3.2 Inference Methods

3.2.1 Exact Inference. The obvious approach for defining the factors used in the CPTG is to perform exact inference based on maximum likelihood estimates (MLE). Specifically, we define the MLE of any combination of medical concepts, C, as

$$P_{MLE}(C) = \frac{\text{number of EMRs in the collection which contain } C \text{ in their CPTs}}{\text{the total number of EMRs in the collection}}, \tag{8}$$

which allows us to define each of the four ϕ factors as

$$\phi_1(\mathcal{D}) = P_{MLE}(\mathcal{D}) \quad \phi_2(\mathcal{S}) = P_{MLE}(\mathcal{S}) \quad \phi_3(\mathcal{E}) = P_{MLE}(\mathcal{E}) \quad \phi_4(\mathcal{R}) = P_{MLE}(\mathcal{R}), \tag{9}$$

and each of the six ψ factors as

$$\begin{aligned} &\psi_1(\mathcal{D},\mathcal{S}) = P_{MLE}(\mathcal{D}\cup\mathcal{S}) \quad \psi_2(\mathcal{S},\mathcal{E}) = P_{MLE}(\mathcal{S}\cup\mathcal{E}) \quad \psi_3(\mathcal{D},\mathcal{E}) = P_{MLE}(\mathcal{D}\cup\mathcal{E}),\\ &\psi_4(\mathcal{D},\mathcal{R}) = P_{MLE}(\mathcal{D}\cup\mathcal{R}) \quad \psi_5(\mathcal{E},\mathcal{R}) = P_{MLE}(\mathcal{E}\cup\mathcal{R}) \quad \psi_6(\mathcal{S},\mathcal{R}) = P_{MLE}(\mathcal{S}\cup\mathcal{R}). \end{aligned} \tag{10}$$

It is important to note that we discover a very large number of CPTs from the EMR collection, and hence the CPTG contains a significantly large number of diagnoses, signs/symptoms, tests, and treatments. Therefore, computing Equation (8) entails either (a) pre-computing $P_{MLE}(C)$ for every possible combination of medical concepts—requiring considering all $2^{|\mathbb{D}\cup\mathbb{S}\cup\mathbb{E}\cup\mathbb{R}|}$ possible combinations—which is prohibitively expensive; or (b) computing $P_{MLE}(C)$ *on-demand*. This alternative approach takes advantage of a bag-of-medical-concepts model. In the bag-of-medical-concepts model, akin to the bag-of-word model used in vector space retrieval, we generated an index that uses as a dictionary of all instances of qualified medical concepts processed from the EMR collection, while the inverted list structures the linked lists of the CPTs that contain each instance of a medical concept. This index is used to compute $P_{MLE}(C)$, when a conjunctive Boolean query is formed with all components of C. The estimations of the factors from Equations (9) and (10) are all produced by using the bag-of-medical-concepts model and the index of CPTs.

3.2.2 Inference with Pair-Wise Smoothing. While the exact inference method accurately estimates the number of EMRs with CPTs containing a given combination of medical concepts, it cannot account for sparsity. For example, if the combination C contains eight medical concepts, although there may be CPTs in the EMR collection that have six or seven medical concepts in common with C, if no CPT in the collection contains all eight medical concepts in C with the same assertions, then the MLE probability will be zero. To address this problem, we relax the maximum likelihood estimates to better handle sparsity. By defining each factor as the product of the pair-wise associations between all concepts C, we are smoothing the MLE of each factor. In this way, the four same-typed factors ($\phi_1 \ldots \phi_4$) can be assigned to the product of pair-wise MLE estimates:

$$\begin{aligned} \phi_1(\mathcal{D}) &= \prod_{d_1\in\mathcal{D}} \prod_{d_2\in\mathcal{D}/\{d_1\}} P_{MLE}(\{d_1,d_2\}), \quad & \phi_2(\mathcal{S}) &= \prod_{s_1\in\mathcal{S}} \prod_{s_2\in\mathcal{S}/\{s_1\}} P_{MLE}(\{s_1,s_2\}),\\ \phi_3(\mathcal{E}) &= \prod_{e_1\in\mathcal{E}} \prod_{e_2\in\mathcal{E}/\{e_1\}} P_{MLE}(\{e_1,e_2\}), \quad & \phi_4(\mathcal{R}) &= \prod_{r_1\in\mathcal{R}} \prod_{r_2\in\mathcal{R}/\{r_1\}} P_{MLE}(\{r_1,r_2\}). \end{aligned} \tag{11}$$

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

$$\begin{aligned} &\psi_1(\mathcal{D},\mathcal{S}) = \prod_{d\in\mathcal{D}}\prod_{s\in\mathcal{S}} P_{MLE}(\{d,s\}), \quad \psi_2(\mathcal{S},\mathcal{E}) = \prod_{s\in\mathcal{S}}\prod_{e\in\mathcal{E}} P_{MLE}(\{s,e\}), \quad \psi_3(\mathcal{D},\mathcal{E}) = \prod_{d\in\mathcal{D}}\prod_{e\in\mathcal{E}} P_{MLE}(\{d,e\}),\\ &\psi_4(\mathcal{D},\mathcal{R}) = \prod_{d\in\mathcal{D}}\prod_{r\in\mathcal{R}} P_{MLE}(\{d,r\}), \quad \psi_5(\mathcal{E},\mathcal{R}) = \prod_{e\in\mathcal{E}}\prod_{r\in\mathcal{R}} P_{MLE}(\{e,r\}), \quad \psi_6(\mathcal{S},\mathcal{R}) = \prod_{s\in\mathcal{S}}\prod_{r\in\mathcal{R}} P_{MLE}(\{s,r\}). \end{aligned} \tag{12}$$

By using the pair-wise definitions from Equations (11) and (12), we estimate the joint distribution in Equation (7).

3.2.3 Inference with Interpolated Smoothing. The pair-wise smoothing method for answer inference still suffers from sparsity problems: if the likelihood of *any pair* of medical concepts is zero, then (as with the exact inference method), the joint probability will be zero. Moreover, the pairwise approach cannot distinguish between the *level of similarity* between a given combination of medial concepts (and their assertions) C and the CPTs used to generated the CPTG. For example, if C contains eight concepts, and there are 50 EMRs that share seven concepts with C but 200 EMRs that share only one concept with C, the 200 EMRs with only a single concept in common would dominate the probability estimates. To account for this, we define the level of similarity between two CPTs as the number of concepts contained in both CPTs (with the same assertions). Thus, the levels of similarity range from perfectly similar (all $|C|$ concepts in common) to perfectly dissimilar (0 concepts in common). To account for each of these levels of similarity, we interpolated the likelihood of C, with the likelihoods of all subsets of C. This would typically require enumerating all $2^{|C|}$ subsets of C and, thus, would appear to be computationally intractable. Fortunately, as with the exact inference method, we can reduce the complexity to be linear in the size of the EMR collection by using the same bag-of-medical-concepts model described in Section 3.2.1.

Specifically, using the same index created for exact inference (described in Section 3.2.1), we were able to compute the smoothed likelihood of a given combination of medical concepts C through a series of constant-time Boolean retrieval operations. Formally, for each medical concept (and its assertion) $c_i \in C$, we construct a separate Boolean query consisting only of c_i and identify the EMRs in the collection that are returned by that query. This allows us to produce a binary vector h_i (for each $c_i \in C$), which indicates which EMRs in the collection mentioned c_i (with the given assertion). We can determined the number of medical concepts in common (i.e., the level of similarity) between each EMR in the collection and C by computing the element-wise sum over each of these binary vectors. We denote the element-wise sum as $m = \sum_i h_i$. Using m, we can compute the number of EMRs in the EMR collection that have each level of similarity with C. Formally, let n_j indicate the number of EMRs in the collection that have a j level of similarity with C. We computed n_j by initializing n as a zero vector and then, for each $m_k \in m$, incrementing n_{m_k} by one. This allows the smoothed likelihood of C to be estimated by interpolating the number of EMRs at each similarity level $(n_0 \ldots n_{|C|})$:

$$P(C) \propto \alpha \cdot n_{|C|} + \sum_{i=1}^{|C|-1} \left[(1-\alpha)^{2^{|C|-i}} \cdot n_i \right], \tag{13}$$

where $\alpha \in [0, 1]$ is a scaling factor that determines how much smoothing is applied: When $\alpha = 1$, no smoothing is applied and Equation (13) reduces to the exact probability estimation (given in Section 3.2.1); when $\alpha = 0$, the exact probability estimation is ignored and only the interpolated similarity counts are used. In our experiments, we used $\alpha = 0.5$.

3.2.4 Bethe Free-Energy Approximation. Finally, we considered state-of-the-art methods for *approximate inference*. In contrast to the three previous approaches for answer inference, approximate inference guarantees a constant upper bounds on the error between the approximate probability and the true probability of each factor. The canonical example of an approximate inference algorithm is that of Loopy Belief Propagation (Pearl 1986), wherein variables and factors repeatedly exchange messages until convergence, at which point, the full joint distribution can be estimated. However, recent work has 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:

$$J(C) = \log \prod_{i=1}^{4} \psi_i(C) \prod_{j=1}^{6} \phi_j(C). \tag{14}$$

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

$$F(C) = U(C) - H(C) = \overbrace{P(C)J(C)}^{\text{energy}} - \overbrace{P(C)\log P(C)}^{\text{entropy}}, \tag{15}$$

where $U(C)$ is the energy and $H(C)$ is the entropy of C. It has been shown the minimum fixed points of Equation (15) are equivalent to fixed points of the iterative Loopy Belief Propagation algorithm, as reported by Vontobel (2013) and Yedidia et al. (2005). This observation indicates that minimizing the free energy in Equation (15) obtains the same probability estimates as running iterative loopy belief propagation on Equation (7) until convergence. We can take advantage of the Bethe free energy approximation by transforming our original, potentially infinitely-looping message passing problem into a convex linear programming problem. As with the pair-wise smoothing approach described in Section 3.2.2, the Bethe free energy approximation relies on pair-wise interactions. Formally,

$$F_B(C,\tau) = U_B(C,\tau) - H_B(C,\tau), \tag{16a}$$

where

$$U_B(C,\tau) = -\sum_{x\in C}\sum_{v_x\in\{0,1\}} \tau_x(v_x)\log\phi(x) \quad - \sum_{y\in C/\{x\}}\sum_{v_y\in\{0,1\}} \tau_{x,y}(v_x,v_y)\log\psi(x,y), \tag{16b}$$

$$H_B(C,\tau) = -\sum_{x\in C}\sum_{v_x\in\{0,1\}} \tau_x(v_x)\log\tau_x(v_x) \quad - \sum_{y\in C/\{x\}}\sum_{v_y\in\{0,1\}} \tau_{x,y}(v_x,v_y)\log\frac{\tau_{x,y}(v_x,v_y)}{\tau_x(v_x)\tau_y(v_y)}. \tag{16c}$$

This allows $P(C)$ from Equation (7) to estimated by finding the set of τ that minimize $F_B(C,\tau)$:

$$P(C) \approx \exp[-\min_{\tau} F_b(C,\tau)], \tag{17a}$$

where τ must satisfy the following conditions:

$$\forall x \in C, v_x \in \{0,1\} \sum_{v_y\in\{0,1\}} \tau_{x,y}(v_x,v_y) = \tau_x(v_x), \tag{17b}$$

$$\forall x \in C, y \in C/\{x\} \sum_{v_x\in\{0,1\}}\sum_{v_y\in\{0,1\}} \tau_{x,y}(v_x,v_y) = 1, \tag{17c}$$

$$\forall x \in X \sum_{v_x\in\{0,1\}} \tau_x(v_x) = 1. \tag{17d}$$

The constraints in Equations (17b)–(17d) can be represented by Lagrangian multipliers, allowing us to estimate the joint probability of any clinical picture and therapy from Equation (7) using gradient descent (or any other method for convex optimization). In our implementation, we used the publicly available Hogwild software for parallel stochastic gradient descent (Recht et al. 2011).

Overall, we have considered four approaches for inferring the probability of any given combination of clinical picture and therapies as defined by Equation (7), enabling us to use Equations (1)–(6) to infer the answers to a medical topic and, consequently, to rank the scientific articles containing each topic. Before computing $P(C)$, however, it is necessary to first construct the CPTG from a collection of EMRs.

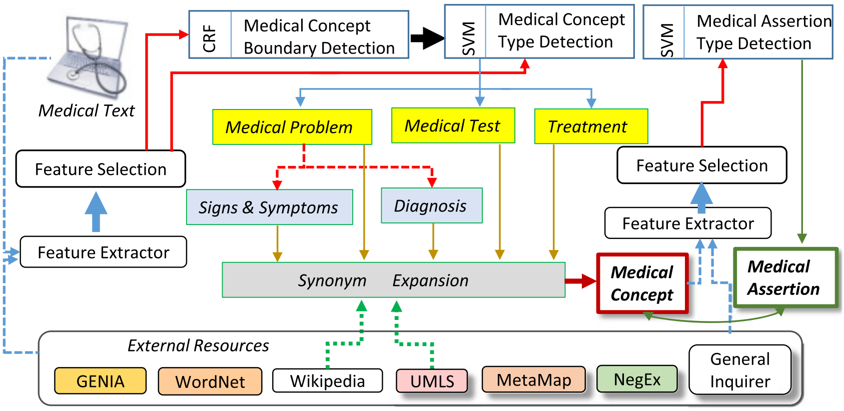

Fig. 6. System for Medical Concept Detection and Assertion Recognition.

4 AUTOMATICALLY GENERATING MEDICAL KNOWLEDGE REPRESENTATIONS WITH MEDICAL LANGUAGE PROCESSING

While the probabilistic inference methods described in Section 3.2 represent the (hyper-)edges in the CPTG as factors, the nodes of the CPTG are identified by applying natural language processing on the collection of electronic medical records (EMRs). The natural language processing extracts the CPT from each medical record by automatically identifying every medical concept (and its assertions) from the natural language narrative. In this section, we detail the natural language processing approach used to automatically identify medical concepts and their assertions. The system that identifies medical concepts and their assertions is illustrated in Figure 6. As will be detailed in Sections 4.1, and 4.2, first medical concepts are identified and then their assertion values are recognized by extracting and selecting features from multiple external resources, as shown in Figure 6. It should be noted that the same natural language processing techniques are applied to the medical topic and scientific articles when producing the medical knowledge sketch. We first describe how medical concepts are discerned, then we detail how assertions are identified.

4.1 Identification of Medical Concepts

When designing our automatic approach to recognizing medical concepts in clinical texts we started from the general framework developed during the 2010 shared-task on *Challenges in Natural Language Processing for Clinical Data* (Uzuner et al. 2011) jointly organized by the Informatics for Integrating Biology at the Bedside (i2b2) and the United States Department of Veteran's Affairs (VA). In this shared-task, participants were asked to identify three types of medical concepts in clinical texts: medical problems, treatments, and medical tests. For our work, we have extended this framework by further classifying medical problems into: (1) observations from the patient (known as SYMPTOMS) or from a physical exam (known as SIGNS); and (2) the DIAGNOSES, including co-morbid diseases or disorders.

We cast the problem of identifying medical concepts in narratives of EMRs as three-stage classification:

Stage 1. Recognizing the boundaries of medical concepts.
Stage 2. Discriminating between medical problems, tests, and treatments.
Stage 3. Classifying medical problems as either SIGNS/SYMPTOMS or DIAGNOSES.

Topic	Diagnoses:	∅
	Signs & Symptoms:	{[**dyspnea**, shortness of breath, sob, . . .]/PRESENT, [**tachypnea**, rapid breathing, increased respiratory rate, . . .]/assertionpresent, [**chest pain**, dthoracic pain, pain in chest, . . .]/PRESENT, [**swollen calf**, calf swelling, swelling of the calf, . . .]/PRESENT, [**pain of right calf**, pain in the right calf, right calf pain, . . .]/PRESENT,}
	Tests:	∅
	Treatments:	∅
Article	**Excerpt:**	. . . *Patients over 75 years, bed rest over four days, cancer, chronic obstructive pulmonary disease, heart failure, kidney failure, tachycardia and syncope are all clinical and comorbidity indicators of poor prognosis in acute PE . . .*
	Diagnoses:	{ [**cancer**, malignant neoplasms, malignant tumor, . . .]/PRESENT, [**chronic obstructive pulmonary disease**, copd, small airway disease, . . .]/PRESENT, [**heart failure**, myocardial failure, cardiac insufficiency, . . .]/PRESENT, [**kidney failure**, renal failure, . . .]/PRESENT, [**acute pulmonary embolism**, acute PE, . . .]/PRESENT}
	Signs & Symptoms	{ [**tachycardia**, rapid pulse, . . .]/PRESENT, [**syncope**, fainting, . . .]/PRESENT,}
	Tests:	∅
	Treatments:	{ [**bed rest**, bedrest, . . .]/PRESENT}

Fig. 7. Example of medical concepts and their assertions discerned from the summary of medical Topic 33 (illustrated in Figure 1) as well as from the relevant PubMed article PMC3913120.[2]

In stage one, a conditional random field (CRF) was used to determine the boundaries (starting and ending tokens) of each medical concept. Stage two of the pipeline relied on a support vector machine (SVM) to determine the type of medical concept. In stage 3, we automatically project each identified medical concept onto the UMLS ontology and classify it as a SIGN/SYMPTOM if the UMLS semantic type is SYMPTOM OR SIGN or FINDING, and as a DIAGNOSIS, otherwise. Stages one and two relied on lexical information, concept type information from UMLS and Wikipedia, as well as semantic information describing predicates and arguments. Automatic feature selection was used to tune the number of features considered by the CRF and SVM separately, as documented in Roberts and Harabagiu (2011). Overall, feature extraction relied on a number of external resources including The Unified Medical Language System (UMLS) (Bodenreider 2004), MetaMap (Aronson 2001), the GENIA project (Kim et al. 2003), WordNet (Fellbaum 1998), and Wikipedia.

In addition to recognizing the boundaries and types of medical concepts, we also associated each medical concept with a set of synonyms (including synonymous abbreviations). Synonyms were generated by (1) collecting all UMLS atoms that share the same concept unique identifier (CUI) and (2) sets of article titles in Wikipedia that all redirect to the same article. This allows us to account for synonymous expressions of the same medical concept in the CPTG by combining all the nodes corresponding to synonymous concepts into a single node representing the set of synonymous concepts. Figure 7 illustrates examples of our approach for automatic medical concept recognition.

4.2 Recognizing the Medical Assertions

Our approach for automatically recognizing assertions for medical concepts extends the framework reported by Roberts and Harabagiu (2011) in which the belief status (or assertion type) of a medical concept is determined by a single SVM classifier. As with medical concepts, we trained our classifier using the annotations produced during the 2010 i2b2/VA challenge (Uzuner et al. 2011). However, in this challenge, only medical problems were annotated with assertions. Consequently, we have extended the assertion values to also apply to tests and treatments as previously reported in Goodwin and Harabagiu (2014). To this end, annotated six new assertion values to describe the degree of the physicians' beliefs for treatments and tests as well. Table 2 lists all 12 assertion values as well as their definitions. In Table 2, the six additional assertion values we annotated are

[2] *Management dilemmas in acute pulmonary embolism*; DOI: 10.1136/thoraxjnl-2013-204667.

Table 2. Assertion Values for Medical Concept in this Table

Assertion Value	Problem	Treatment	Test	Definition
PRESENT	✓			The indicated problem is still active at this moment.
ABSENT	✓	✓	✓	The medical concept does not exist at this moment.
POSSIBLE	✓			The patient may have a problem, but there is uncertainty.
HYPOTHETICAL	✓			The patient may develop the indicated problem.
CONDITIONAL	✓	✓	✓	The medical concept occurs only during certain conditions.
ASSOCIATED WITH ANOTHER	✓			The medical problem is associated with someone other than the patient.
⋆ORDERED			✓	The indicated treatment will be completed in the immediate future.
⋆PRESCRIBED		✓		The indicated treatment will begin sometime after this moment.
⋆CONDUCTED			✓	The indicated medical test been performed.
⋆ONGOING	✓	✓		The indicated problem or treatment persists beyond this moment.
⋆SUGGESTED		✓		The indicated treatment or test is advised but is not certain to occur.
⋆HISTORICAL	✓	✓	✓	The indicated medical concept occurred during a previous hospital visit.

In this Table a "moment" Refers to the Specific Instant in Time in which the Particular Medical Concept was Written.

indicated with a "⋆." As previously reported in Goodwin and Harabagiu (2014), we have annotated a total 2,349 medical concepts with the expended set of assertion values.

The SVM used for automatic assertion classification considered the same set of features and external resources as those reported in Roberts and Harabagiu (2011), namely, UMLS, MetaMap, NegEx (Chapman et al. 2001) and the Harvard General Inquirer (Stone et al. 1966). As in Roberts and Harabagiu (2011), we relied on (1) the above external resources, (2) lexical features, and (3) statistical information about assertions classified for previous mentions of the same medical concept to train a 12-class SVM.

5 EXPERIMENTAL RESULTS

In our experiments, we evaluated (1) the accuracy of answers produced by our system for each topic, (2) the relevance of scientific articles retrieved for each topic, and (3) the structure and composition of the automatically generated clinical picture and therapy graph (CPTG). When evaluating the answers and scientific articles automatically produced by our system, we considered the 30 topics (labeled 31–60) used during the 2015 TREC-CDS evaluation (Simpson et al. 2014) (for which answers had been given).

5.1 Medical Answer Evaluation

The accuracy of the medical answers automatically produced by our approach was measured by computing the Mean Reciprocal Rank (MRR). The MRR is the mean of the reciprocal of the rank produced by our system for the (first) correct answer for each topic. To identify the correct answer for each topic, we relied on a set of "candidate answers" manually produced by the authors of the 2015 TREC-CDS topics. The candidate answers were distributed to the TREC-CDS participants after the evaluation had completed. It should be noted that these candidate answers were not provided to the relevance assessors when evaluating document retrieval, and they were only provided to participating teams after the evaluation had concluded. The 2015 TREC-CDS task was strictly an information retrieval evaluation measuring only the performance of systems when retrieving and rank scientific articles from the PubMed Central open access subset. However, the candidate

Table 3. The Mean Reciprocal Rank (MRR) of the Medical Answers Inferred Using Each Inference Method and Each Type of Medical Knowledge Sketch

Inference Method	**Case 1** (Z_1, RLA_1)	$\star$ **Case 2** (Z_2, RLA_2)	**Case 3** (Z_3, RLA_3)
Exact Inference	0.031	0.000	0.220
Pair-wise Smoothing	0.083	0.502	0.329
Interpolated Smoothing	0.124	0.601	0.466
$\star$Bethe Approximation	0.125	0.694	0.464

answers provided after the conclusion of the evaluation allowed us to cast the TREC-CDS task as a question-answering (Q/A) problem. The candidates answers produced by the topic creators indicate one-or-more candidate answers that the topic author considered when producing each topic. It is important to note that the candidate answers are not necessarily the "best" answers and are not always well-represented in PubMed. We evaluated each of the three ranked list of answers produced by our system using the candidate answers as a gold-standard. Table 3 lists the performance of our approach when considering each type of medical knowledge sketch (as described in Section 2) and when using each method of answer inference (as described in Section 3).

As shown in Table 3, it is clear that the most accurate answers were obtained when using (1) the Bethe Free-Energy Approximation method for answer inference applied to (2) the medical knowledge sketch obtained by considering both the medical topic t and an entire scientific article l, $z_2(t, l)$. Clearly, exact inference was unable to produce accurate answers, as evidence by the low performance across all three types of medical knowledge sketch. When investigating the poor performance when $Z_2(t, l)$ was used with exact inference, we found that very few CPTs in our EMR collection contained all the medical concepts with the same assertions as in Z_2, highlighting the high degree of sparsity in clinical data. It is important to point out that the was no statistically significant difference between the Bethe Free-Energy Approximation and the Interpolated Smoothing methods of answer inference (using the Wilcoxon signed-ranked test, where $p < 0.001$ and $N = 30$), but that both methods did significantly outperform both the pair-wise smoothing and exact inference methods of answer inference. This suggests that the accuracy of answers inferred using the CPTG can be greatly improved by smoothing or approximation (rather than using exact estimates). Overall, the answers obtained using $Z_1(t)$ were of significantly poorer quality than those obtained using either $Z_2(t, l)$ or $Z_3(t, s)$. Moreover, the answers obtained using $Z_2(t, s)$ were less accurate than those obtained using $Z_3(t, l)$. While considering paragraphs from scientific articles can improve the accuracy of answers automatically identified by our approach, decrease in performance from $Z_2(t, l)$ to $Z_3(t, s)$ suggests that the most accurate answers (and their rankings) are achieved by considering the entire scientific article. Moreover, the high performance of $Z_2(t, s)$ suggests that the questions associated with medical topics are complex: the answer(s) are rarely described within a single paragraph and must be inferred from multiple paragraphs across the document. We analyzed the answers obtained using each type of medical knowledge sketch and found that the answers produced by considering $Z_1(t)$ were dominated by the most common diseases, tests, or treatments mentioned in the EMR collection (that is, $Z_1(t)$ effectively reduced to the prior probability of each medical concept). By contrast, $Z_2(t, l)$ was able to identify reasonably accurate answers in most cases, while $Z_3(t, s)$ preferred answers that were related to only a small number of medical concepts in the topic, and were often unable to account for the fact that some concepts in the medical topic are more important than others when determining an answer. Overall, the performance measures reported in Table 3 reinforce our hypothesis that combining knowledge from relevance scientific articles with the medical topic can produce significantly more accurate

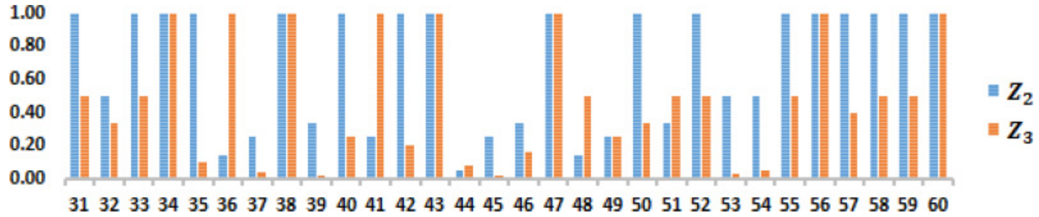

Fig. 8. Reciprocal Rank for each topic evaluated in TREC-CDS 2015.

Table 4. Examples of Answers Discovered for the Medical Cases Illustrated in Figure 1

Topic	Details
33	**EMAT:** Diagnosis **Answers:** acute pulonary embolism; thrombolysis; ischaemia; dvt; pulmonary hypertension; myocardial infarction; tension pneumothorax; arrhythmia; cardiogenic shock; aortic dissection **Gold Answer:** pulmonary embolism
42	**EMAT:** Test **Diagnosis:** bacterial meningitis **Answers:** spinal puncture; gram's stain; cerebrospinal fluid culture; latex fixation test; bacterial cultures; cervical puncture; CSF assessment; lymph node biopsy; cranial nerve assessment; computer tomographic angiogram **Gold Answer:** spinal tap / cerebrospinal fluid analysis
54	**EMAT:** Treatment **Diagnosis:** community acquired pneumonia (cap) **Answers:** continuous positive airway pressure; antibiotics; moxifloxacin; fentanyl; levofloxacin; zosyn; vancomycin; fluid resuscitation; combination therapy **Gold Answers:** antibiotics

answers when considering the topic in isolation. Moreover, the MRR obtained when considering $Z_2(t,l)$ and $Z_3(t,s)$ show that answer inference over the clinical picture and therapy graph are reasonably accurate compared to the candidate answers produced by the 2015 TREC-CDS topic creators.

In addition to the Mean Reciprocal Rank shown in Table 3, Figure 8 shows the reciprocal rank of the gold-standard answer for each individual topic used in the 2015 TREC-CDS evaluation when using the Interpolated Smoothing method for answer inference applied to (1) $Z_2(t,l)$ and (2) $Z_3(t,s)$. As shown, for the majority of topics, for all medical knowledge sketch types, our top-ranked answer was the same (or synonymous with) the candidate answer produced by the topic creators. In fact, we obtained the correct answer for the majority of topics with an EMAT of treatment (i.e., topics 51–60) as well as for many of the topics with an EMAT of diagnosis (i.e., topics 31–40). Unfortunately, for many of the topics with an EMAT of medical test (i.e., topics 41–50), our approach struggled to identify and rank the correct answers. Table 4 presents the ten highest-ranked answers produced by our approach for each topic previously shown in Figure 1 as well as the (1) the candidate answer(s) produced by the TREC topic authors and (2) the held-out diagnosis.

As shown, for Topics 33 and 42, we obtain the correct answer at the highest rank, while for Topic 54, we obtain the correct answer at the second rank. In the case of Topic 33, this is because *pulmonary embolisms* were mentioned with high frequency in relevant scientific articles and were often diagnosed for patients with many of the signs/symptoms and tests as indicated by the topic. The answers obtained at lower ranks include other cardiovascular conditions such as deep vein thrombosis (*DVT*) and heart attack (*myocardial infarcation*), with a handful of less severe co-morbid conditions such as low blood flow (*ischaemia*) and high blood pressure (*hypertension*). When answering Topic 42, our system produced a synonym *spinal puncture* for the correct answer of *spinal tap*. The second-ranked answer, *gram's stain* is a test used to distinguish between types of bacteria in a collected sample or culture. Note that cerebrospinal fluid (CSF) culture and CSF assessment from the correct answer were obtained at ranks three and seven. This indicates that although our method was able to identify the most important medical concept for treating bacterial meningitis, it was not as effective at ranking other highly related concepts. It should be noted that Topic 42 was one of only three topics with an expected medical answer type (EMAT) of Test in which the correct answer was produced at rank one. For the majority of test topics, the most commonly described tests related to the topic were ranked at the highest positions, rather than the tests most related to the topic. Finally, for Topic 54, we obtained the correct answer *antibiotics* at the second rank, followed by a large number of specific antibiotics (e.g., *mxofloxacin, fentanyl, levofloxacin*, etc.). Interestingly, the highest-ranked treatment for Topic 54 was *continuous positive airway pressure*, or CPAP. Although CPAP is a treatment used to keep the airways open for patients with respiratory problems, it only treats the symptoms rather than the underlying pneumonia. This suggests a possible area for future research: understanding which treatments target the disease itself (such as antibiotics) and which treatments are designed to alleviate symptoms of the disease (such as CPAP).

Overall, we found two main sources of errors in the answers produced by our system: (1) a failure to account for more fine-grained relationships between individual tests and treatments and the specific medical problems the treatments are targeting, as well as (2) the inability to account for *counter-indicated* medical treatments—medical treatments that are known to produce adverse effects in certain situations. Unlike electronic medical records (EMRs), many scientific articles begin with a review of recent literature and the current knowledge of their subject, often presenting negative findings or counter-indicated treatments, such as medications that are typically prescribed for a disease but are known to produce adverse reactions in specific situations. This type of counter-indicated information is not negated, or speculated, and suggests the need for identifying either more fine-grained assertions, or detecting counter-indication relations in scientific articles and electronic medical records.

5.2 Medical Article Retrieval Evaluation

The performance of our approach when automatically identifying and ranking scientific articles relevant to each medical topic was measured by relying on the relevance judgments produced during the 2015 TREC-CDS topics by Oregon Health and Science University (OHSU). For the 2015 topics, a total of 37,807 topic-article pairs were judged as either (1) relevant, (2) partially relevant, or (3) non-relevant. Physician students provided relevance judgments by pooling the 20 top-ranked articles as well as a 20% random sample of the articles retrieved between ranks 21 and 100 for each topic retrieved by any participating team in the TREC CDS task.

We followed the official TREC-CDS evaluations, by not distinguishing between relevant and partially relevant articles when measuring the performance of our system (i.e., we considered only binary relevance). We measured the quality of the ranked list of scientific articles produced by our approach using four information retrieval metrics also used in TREC: (1) the inferred Average Precision (inf. AP), wherein retrieved articles were randomly sampled and the Average Precision was calculated as in Yilmaz and Aslam (2006); (2) the inferred Normalized Discounted Cumulative Gain (inf. NDCG), wherein retrieved articles were randomly sampled and the NDCG was calculated as per (Yilmaz et al. 2008); (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) (Manning et al. 2008).

We compared the quality of ranked scientific articles produced by our system using each answer inference method applied to each type of medical knowledge sketch and found that the best performance was obtained when using Interpolated Smoothing for answer inference and $Z_2(t, l)$

Table 5. Performance Results Obtained for the System Reported in This Paper (Q/A-CDS) When Using Each Type of Medical Knowledge 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

	inf. AP	**inf. NDCG**	***R*-Prec**	**P@10**
⋆ Baseline: BM25	.042	.204	.163	.387
Baseline: TF-IDF	.041	.197	.169	.350
Baseline: LMJM	.040	.193	.151	.357
Baseline: LMDir	.043	.203	.170	.360
Baseline: DFR	.039	.197	.167	.333
$Z_1(t)$	.006	.010	.020	.062
⋆$Z_2(t, l)$	.147	.434	.344	.722
⋆$Z_3(t, s)$	.018	.114	.081	.190
Exact Inference	.063	.167	.154	.410
Pair-wise Smoothing	.128	.382	.330	.610
⋆Interpolated Smoothing	**.147**	**.434**	**.344**	**.722**
Bethe Approximation	.140	.432	.336	.701

as the medical knowledge sketch. Consequently, for readability, in Table 5, we present the results when (1) using interpolated smoothing for answer inference on each type of medical knowledge sketch, as well as (2) when using each type of answer inference for $Z_2(t, l)$. Moreover, Table 5 also indicates the performance obtained using five baseline information retrieval models: BM25 relied on the Okapi-BM25 (Robertson et al. 1995) ($k_1 = 1.2$ 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 (Zhai and Lafferty 2001), respectively; and DFR considered the Divergence from Randomness framework (Amati and Van Rijsbergen 2002) 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 Tasks A and B. Note that, as in the official evaluation, we distinguish between automatic systems that involved no human intervention and manual systems in which arbitrary human intervention was allowed. For Task A (in which no explicit diagnoses was provided), the best inf. NDCG reported for an automatic system was 0.294 and the best for a manual system was 0.311, while for Task B (in which an explicit diagnoses was given for each topic focusing on a medical test and treatment), the best inf. NDCG reported for an automatic system was 0.382 and the best reported for a manual system was 0.381. Please note that our approach was designed for Task A and, consequently, does not consider the explicit diagnoses given in Task B. Moreover, our approach incorporates a very basic form of query expansion (described in Section 3.1) and a simple relevance model (BM25) while many of 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) that were not considered in our approach (Roberts et al. 2016).

Clearly, the best performance obtained by our system (denoted with a "⋆") relies on (1) the medical knowledge sketch obtained by consider the topic t and scientific article l ($Z_2(t, l)$) and (2) the interpolated-smoothing method for answer inference. Likewise, we found that the best relevance model for our approach (based on the performance of each baseline) was the BM25 ranking function. There was no statistically significant difference in the performance measured when applying the Interpolated Smoothing or the Bethe Approximation methods for answer inference (as observed when measuring the accuracy of answers produced by our approach). Clearly, as illustrated in Table 5, our approach enabled significantly higher quality scientific article retrieval than the top reported systems for each task (Simpson et al. 2014). Specifically, we measured a 49% increase in inferred NDCG compared to the best reported automatic system (Balaneshin-kordan et al. 2015) and measured a 40% increase in inferred NDCG compared to the best reported manual system (Balaneshin-kordan et al. 2015) when considering Task A. When comparing our approach to Task B, in which an explicit diagnosis was provided with every topic with an EMAT of test or treatment, we measured a 14% increase in inferred NDCG compared to the best reported automatic (Song et al. 2015) and manual (You et al. 2015) systems. We believe that the difference in performance increase observed between our approach and the state-of-the-art approaches across task A and B indicates that our approach was often able to infer the correct diagnosis in Task A. Moreover, we also observed a clear increase in performance when comparing our approach to the state-of-the-art for Task B, which suggests that the ability of our approach to infer additional semantically meaningful medical concepts beyond the explicit diagnosis that were able to improve the relevance of retrieved scientific articles. This in turn, further suggests that the relevant articles in the TREC-CDS task contain answers that were not always in the candidate answer set. Overall, we believe that the high performance of our approach clearly demonstrates the impact of incorporating medical question answering from knowledge bases to improve clinical decision support.

5.3 Medical Knowledge Evaluation

There is no clear way to measure the "accuracy" of the clinical picture and therapy graph (CPTG) (as the edges between concepts do not indicate a single, direct semantic relationship). Consequently, we report the structure and connectivity of the CPTG. The CPTG contained 634 thousand nodes and 13.9 billion edges where 31.2% of all nodes were diagnoses, 21.84% were signs or symptoms, 23.62% were medical tests, and 23.34% of nodes were medical treatments. The distribution of assertions associated with medical concepts in the CPTG is: 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. To evaluate the CPTG, we evaluated only the quality of the nodes of the CPTG that represent medical concepts and their assertions. We were unable to evaluate the edges of the CPTG, because there are no medical knowledge bases or ontologies that encode relations between medical concepts qualified by assertions.

The evaluation of the quality of the nodes encoded in the CPTG considered the F_1-scores when (1) detecting the boundaries of medical concepts; (2) detecting the type of medical concepts; and (3) identifying assertions. To perform the evaluation, we considered the $72,846$ gold-standard annotations provided with the 2010 i2b2/VA shared-task. On that data, our system, as reported in Roberts and Harabagiu (2011), obtained an F_1-score of 83.45% for boundary detection, 95.49% for concept type detection, and 93.94% for assertion recognition. However, as noted in Section 4, the i2b2 annotations did not indicate whether medical problems were signs/symptoms or diagnoses and did not include assertions for medical tests or treatments. Consequently, we performed $2,349$ additional annotations on EMRs from MIMIC III. We performed 10-fold cross validation, which allowed us to compute the F_1-scores when detecting medical concept boundaries as 81.22%, whereas the F_1-score when detecting medical concept types was 85.99% and the F_1-score for identifying assertions was 75.99%. It is obvious that the new concept types and assertions that were annotated impacted the performance of the automatic medical concept and assertion identification system. We believe that the performance may be improved as more annotations become available for training.

6 CONCLUSIONS

In this article, we detail the knowledge representations considered in a novel medical Q/A framework that can be used in CDS systems for recognizing relevant biomedical articles and pinpointing the answers to questions about complex medical cases. To answer medical questions about complex medical cases, we introduced the notion of medical knowledge sketches that capture the clinical background of the medical case. We have presented three forms of medical knowledge sketches and shown how they can be used to infer answers. Moreover, we have shown how four different probabilistic inference methods operate on the medical knowledge acquired from a vast EMR collection, reflecting knowledge pertaining to medical practice. We also introduced three novel article relevance models, informed by answers, which are used to retrieve relevant biomedical articles.

In our experiments, we considered all 12 combinations of medical knowledge sketches and probabilistic inference methods and the results indicated surprisingly high MRR scores of the answers 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. The best results were obtained when the medical knowledge sketch considered all the medical knowledge discerned from an entire biomedical article as well as the medical knowledge discerned from the description of the medical case, while using the Interpolated smoothing method of probabilistic inference (Bethe approximation performed equally well). Moreover, when the answers of a medical question are known, they inform the ranking of relevant articles from PubMed with 86.5% increased inferred Average Precision to current state-of-the-art systems evaluated in the most recent TREC-CDS.

To our knowledge this is the first work that considers medical Q/A from a knowledge base by employing four probabilistic inference methods. It is also the first attempt of representing knowledge as (a) medical knowledge sketches and (b) a clinical picture and therapy graph. Possible avenues for future work include (1) automatically recognizing semantic attributes such as severity, temporality, and so on, and incorporating them into the knowledge graph, and (2) considering the roles of different sections, rather than paragraphs or entire articles when constructing the medical knowledge sketch.

REFERENCES

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

Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *Proceedings of the American Medical Informatics Association (AMIA'01)*. 17.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. *Dbpedia: A Nucleus for a Web of Open Data*. Springer.

Saeid Balaneshin-kordan, Alexander Kotov, and Railan Xisto. 2015. WSU-IR at TREC 2015 clinical decision support track: Joint weighting of explicit and latent medical query concepts from diverse sources. In *Proceedings of the Text Retrieval Conference (TREC'15)*.

Junwei Bao, Nan Duan, Ming Zhou, and Tiejun Zhao. 2014. Knowledge-based question answering as machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 967–976.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* 32, suppl. 1 (2004), D267–D270.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the International Conference on Management of Data (SIGMOD'08)*. ACM, 1247–1250.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F. Cooper, and Bruce G. Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics (JBI)* 34, 5 (2001), 301–310.

Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 2015. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the Association for Computational Linguistics (ACL'15)*, Vol. 1. 260–269.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Amit X. Garg, Neill K. J. Adhikari, Heather McDonald, M. Patricia Rosas-Arellano, P. J. Devereaux, Joseph Beyene, Justina Sam, and R. Brian Haynes. 2005. 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 (2005), 1223–1238.

Travis Goodwin and Sanda M. Harabagiu. 2014. Clinical data-driven probabilistic graph processing. In *Proceedings of the Conference on Language Resources and Evaluation (LREC'14)*. 101–108.

Travis R. Goodwin and Sanda M. Harabagiu. 2016. Medical question answering for clinical decision support. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 297–306.

Sanda Harabagiu and Steven Maiorano. 1999. Finding answers in large collections of texts: Paragraph indexing + abductive inference. In *Proceedings of the AAAI Fall Symposium on Question Answering Systems*. 63–71.

J.-D. Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus—A semantically annotated corpus for bio-textmining. *Bioinformatics* 19, suppl. 1 (2003), i180–i182.

Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.

Joon Lee, Daniel J. Scott, Mauricio Villarroel, Gari D. Clifford, Mohammed Saeed, and Roger G. Mark. 2011. Open-access MIMIC-II database for intensive care research. In *Proceedings of the Conference on Engineering in Medicine and Biology Society (EMBC'11)*. IEEE, 8315–8318.

Carolyn E. Lipscomb. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* 88, 3 (2000), 265.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*, Vol. 1. Cambridge University Press, Cambridge.

Judea Pearl. 1986. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence* 29, 3 (1986), 241–288.

Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. 2011. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS'11)*. 693–701.

Kirk Roberts and Sanda M. Harabagiu. 2011. A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association (JAMIA)* 18, 5 (2011), 568–573.

Kirk Roberts, Matthew Simpson, Dina Demner-Fushman, Ellen Voorhees, and William Hersh. 2016. State-of-the-art in biomedical literature retrieval for clinical cases: A survey of the TREC 2014 CDS track. *Info. Retriev. J.* 19, 1–2 (2016), 113–148.

Kirk Roberts, Matthew S. Simpson, Ellen Voorhees, and William R. Hersh. 2015. Overview of the TREC 2015 clinical decision support track. In *Proceedings of the 24th Text Retrieval Conference Proceedings (TREC'15)*.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, Mike Gatford et al. 1995. Overview of the Third Text REtrieval Conference (TREC-3). National Institute of Standards and Technology (NIST). 109–126.

Richard H. Scheuermann, Werner Ceusters, and Barry Smith. 2009. Toward an ontological treatment of disease and diagnosis. American Medical Informatics Association (AMIA). 116–120.

Matthew S. Simpson, E. Voorhees, and William Hersh. 2014. Overview of the TREC 2014 clinical decision support track. In *Proceedings of the Text Retrieval Conference (TREC'14)*. National Institute of Standards and Technology.

Yang Song, Yun He, Qinmin Hu, and Liang He. 2015. ECNU at 2015 CDS track: Two re-ranking methods in medical information retrieval. In *Proceedings of the 2015 Text Retrieval Conference (TREC'15)*.

Philip J. Stone, Dexter C. Dunphy, and Marshall S. Smith. 1966. The general inquirer: A computer approach to content analysis. *Journal of Regional Science* 8, 1 (1966), 113–116.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association (JAMIA)* 18, 5 (2011), 552–556.

Harold Varmus, D. Lipman, and P. Brown. 1999. PubMed central: An NIH-operated site for electronic distribution of life sciences research reports. National Institutes of Health (NIH). 24 (1999), 1999.

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

William A. Woods. 1973. Progress in natural language understanding: An application to lunar geology. In *Proceedings of the June 4–8, 1973, National Computer Conference and Exposition*. ACM, 441–450.

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *Proceedings of the Association for Computational Linguistics (ACL'14)*. Citeseer, 956–966.

Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. 2005. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory* 51, 7 (2005), 2282–2312.

Emine Yilmaz and Javed A. Aslam. 2006. Estimating average precision with incomplete and imperfect judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. ACM, 102–111.

Emine Yilmaz, Evangelos Kanoulas, and Javed A Aslam. 2008. A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 603–610.

Ronghui You, Yuanjie Zhou, Shengwen Peng, Shanfeng Zhu, and R. China. 2015. FDUMedSearch at TREC 2015 clinical decision support track. In *Proceedings of the Text Retrieval Conference (TREC'15)*.

Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the Special Interest Group on Information Retrieval (SIGIR'01)*. ACM, 334–342.

Received January 2017; revised April 2017; accepted May 2017