Medical Question Answering for Clinical Decision Support

Travis R. Goodwin and Sanda M. Harabagiu
The University of Texas at Dallas
Human Language Technology Research Institute
http://www.hlt.utdallas.edu
What is Clinical Decision Support (CDS)?

“Clinical decision support (CDS) provides clinicians, staff, patients or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care.” – Healthit.gov

In our case, anticipate the needs of physicians

- link **electronic medical records (EMRs)** with relevant **bio-medical literature**.
Special track on **Clinical Decision Support** in the **Text REtrieval Conference (TREC-CDS)**

Given a **medical case**, answer one of three generic medical questions:
1. *What is the diagnosis?*
2. *What test(s) should be ordered?*
3. *Which treatment(s) should be administered?*

“**Topic**” = Medical case (description & summary) + Expected Medical Answer Type (EMAT)
Goal: Given a topic, return a ranked list of scientific articles from PubMed Central containing the answers.

- IR Evaluation – NOT Q/A
- In 2015, a new track in which the diagnosis was explicitly provided
- Performance when the diagnoses were given was significantly improved.

Research questions:
- Could we automatically identify this type of information, i.e. “partial answers” to each topic?
- Rather than retrieving documents, and finding the answer within those documents, would it be better to reverse the process?
Standard TREC-CDS System

TREC-CDS Topic → Topic Processing → Query Expansion

PubMed Articles → Index → Ranking → Ranked List of Relevant Articles
Introduction

Cast the problem of retrieving scientific articles for medical topics as a question answering (Q&A) from knowledge base(s) problem.

Our approach:

1. Generate a large medical knowledge graph from publicly available EMRs
2. Apply probabilistic inference on the knowledge graph to determine the answer to a topic
3. Use the automatically determined answer(s) to retrieve and rank scientific articles from PubMed
Presentation Outline

1. Introduction
2. System Architecture
3. Inferring the Answers
4. Experimental Results
5. Conclusions
Generating the Knowledge Graph

Definitions:

- **Clinical Picture**: set of all medical problems, signs, symptoms, and tests that might influence the diagnosis of a patient
- **Therapy**: set of all treatments, cures, and preventions included within the management plan of a patient.

The **Clinical Picture and Therapy Graph (CPTG)** captures the variations in clinical picture and therapies (CPTs) from the MIMIC-III patient population.
- **Factorized Markov Network** measuring the likelihood of any possible CPT
- 634 thousand **nodes** corresponding to a medical diagnosis, sign/symptom, test, or treatment
- 14 billion **edges** measuring the cohesive strength between nodes in MIMIC EMRs
The Clinical Picture and Therapy Graph

Consider each node (i.e. concept) as a binary random variable:
- 1, if assertion is PRESENT, SUGGESTED, ORDERED, PRESCRIBED
- 0, if assertion is ABSENT
- considered a latent variable, otherwise.

Represent edges between nodes as factors:
- 6 factors between sets of concepts of different types:
  - e.g., $\psi_1(\mathbb{D}, \mathbb{S})$: likelihood that a CPT contains the diagnoses in $\mathbb{D}$ as well as the signs/symptoms in $\mathbb{S}$
- 4 factors for sets of concepts with the same type:
  - e.g., $\phi_1(\mathbb{D})$: (prior) likelihood that a CPT contains all the diagnoses in $\mathbb{D}$
The CPTG can measure the probability of any possible combination \( \mathcal{C} \) of:
- diagnoses \( \mathcal{D} \subseteq \mathcal{C} \)
- signs/symptoms \( \mathcal{S} \subseteq \mathcal{C} \)
- tests \( \mathcal{E} \subseteq \mathcal{C} \)
- treatments \( \mathcal{R} \subseteq \mathcal{C} \)

(Medical sketch is a combination!)

where:
- \( \mathcal{D} \) is the set of all diagnoses\(^1\)
- \( \mathcal{S} \) is the set of all signs/symptoms\(^1\)
- \( \mathcal{E} \) is the set of all tests\(^1\)
- \( \mathcal{R} \) is the set of all treatments\(^1\)

\(^1\) observed in the MIMIC-III collection
Joint distribution over all possible CPTS is computed using the CPTG:

$$P(C) \propto \prod_{i=1}^{6} \psi_i(C) \times \prod_{j=1}^{4} \phi_j(C)$$

How do we compute $\psi_i$ and $\phi_j$?

Use maximum likelihood estimations (MLE)

• $\psi_i(X, Y) = \text{MLE}(X \cup Y)$
• $\phi_j(X) = \text{MLE}(X)$
• $\text{MLE}(W) = \frac{\text{the number of EMRs with } W \text{ in their CPT}}{\text{the total number of EMRs}}$
The Medical Sketch

Consider the medical topic as an “incomplete sketch” of the CPT of a patient

The idea:
Determine the answer by finding the medical concept which matches the expected medical answer type (EMAT) which is most likely to “complete” the CPT.

Case 1:
• Only consider the medical topic
• Represent the combination of medical concepts in the medical topic as a sketch of the patient’s CTP
• Denote the sketch as $Z(t)$

Case 2:
• Combine the medical topic with relevant scientific articles
• Represent the combination of medical concepts in the medical topic as well as those of a single relevant scientific article
• Denote this extended sketch as $EZ(t, l)$
Selecting Answers

In both cases, $z \in \{Z(t), EZ(t, l)\}$, the answer to a medical topic can be determined by:

$$\hat{a} = \arg \max_{a \in A} P(a \cup z) = \frac{P(a \cup z)}{P(z)}$$

- $A$ is the set of all medical concepts matching the EMAT.
- Probability is measured by the CPTG

What about Case 2?
- Consider the relevance of article $l$
- “Reciprocal-Rank Conditional Score” (RRCS):

$$RRCS(a) = \sum_{r=1}^{[L]} \frac{1}{r} \times \frac{P(a \cup EZ(t, l_r))}{P(EZ(t, l_r))}$$
Ranking Scientific Articles

Determine the relevance of an article $l_i$ based on the probability of all answers it contains, $Y_i$

**Case 1:** $\text{Rel}(l_i) \propto P(Y_i \cup Z(t))$

**Case 2:** $\text{Rel}(l_i) \propto \frac{P(EZ(t,l_i))}{P(EZ(t,l_i)-Y_i)}$
Presentation Outline

1. Introduction
2. System Architecture
3. Inferring the Answers
4. Experimental Results
5. Conclusions
Answer Inference

Directly computing the joint distribution over all possible clinical pictures and therapies is not easy!

– $2^{D+S+E+R}$ possible CPTs
– Unlikely that a patient in MIMIC had the exact same CPT as the medical sketch or extended medical sketch

We considered 3 inference methods for estimating the joint distribution:

1. Approximate inference based on Bethe free-energy
2. Pair-wise variational inference
3. Interpolated smoothing
Method 1: Free-Energy Setting

• Provides the same solution as loopy belief propagation, but can be solved with linear programming.

• Cast the joint distribution as the energy in a physical system:

\[
J(C) = \log \prod_{i=1}^{6} \psi_i(C) \times \prod_{j=1}^{4} \phi_j(C)
\]

• We minimize the “Free Energy” of the system:

\[
F(C) = U(C) - H(C) = P(C)J(C) - P(C) \log P(C)
\]
**Method 1: Bethe Free-Energy Approximation**

Bethe Free Energy Approximation:

\[ P(\mathcal{C}) \approx \exp \left[ - \min_{\tau} F_b(\mathcal{C}, \tau) \right] \]

where

\[ F_b(\mathcal{C}, \tau) = U_b(\mathcal{C}, \tau) - H_b(\mathcal{C}, \tau) \]

\[ U_b(\mathcal{C}, \tau) = - \sum_{x \in \mathcal{C}} \sum_{v_x \in \{0,1\}} \tau_x(v_x) \log \phi(x) - \sum_{y \in \mathcal{C}/\{x\}} \sum_{v_y \in \{0,1\}} \tau_{x,y}(v_x, v_y) \log \psi(x, y) \]

\[ H_b(\mathcal{C}, \tau) = - \sum_{x \in \mathcal{C}} \sum_{v_x \in \{0,1\}} \tau_x(v_x) \log \tau_x(v_x) - \sum_{y \in \mathcal{C}/\{x\}} \sum_{v_y \in \{0,1\}} \tau_{x,y}(v_x, v_y) \log \tau_{x,y}(v_x, v_y) \]
Method 2: Pair-wise Variational Inference

Bethe Approximation is very powerful, but is it necessary?

A much simpler alternative:
• “Pair-wise” Variational Inference
• Define each factor as the product of all pair-wise probabilities:

\[
\phi_j(C) = \prod_{x \in C} \prod_{y \in C / \{x\}} \text{MLE} \left( \{x, y\} \right)
\]

\[
\psi_i(C) = \prod_{x \in C} \prod_{y \in C / \{x\}} \text{MLE} \left( \{x, y\} \right)
\]

• Effectively solves “sparsity” problems
Method 3: Interpolated Smoothing

Pair-wise probabilities are too “noisy”

- Do not consider the *similarity* between a given CPT (e.g. \( \mathcal{C} \)) and the CPTs used to generate the CPTG

**Idea:** modify the MLE estimates based on the “*level of similarity*” between each CPT used to build the CPTG and \( \mathcal{C} \):

- “Level of Similarity”: the number of *medical concepts* in both the given CPT (e.g. \( \mathcal{C} \)) and one of the CPTs in the CPTG
- 0 if no medical concepts overlap between a CPT and \( \mathcal{C} \)
- \(|\mathcal{C}|\) if all medical concepts in \( \mathcal{C} \) are in the CPT
Method 3: Interpolated Smoothing II

How do we determine the number of CPTs at each level of similarity?

- Use an inverted index!
- Represent each CPT as a bag-of-words vector
- The vocabulary is the set of all medical concepts in the EMR collection

Four steps:

1. For each concept $c \in C$, retrieve a binary vector with an entry for each EMR, indicating whether $c$ was
   - PRESENT, SUGGESTED, ORDERED, PRESCRIBED with value 1
   - 0, otherwise
2. Let $m$ be the element-wise sum of these binary vectors retrieved for each $c \in C$
   - $m_i = \text{number of concepts in common between CPT}_i$ and $C$
3. Let $n$ by a $|C|$-length vector indicating the number of CPTs at each level of similarity
   - Compute in a single pass over $m$
4. Compute the interpolated probability:

$$P(C) \propto \alpha \cdot n_{|C|} + \sum_{i=1}^{|C|-1} (1 - \alpha)^{2|C|-i} \cdot n_i$$
Experimental Results: Answers

- Used the “candidate” answers provided after the conclusion of TREC-CDS 2015.
- Computed the **Mean Reciprocal Rank (MRR)** using each inference method for both medical sketches:

<table>
<thead>
<tr>
<th>Method</th>
<th>$Z(t)$</th>
<th>$\star E Z(t,l)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>★Bethe Approximation</td>
<td>0.125</td>
<td>0.694</td>
</tr>
<tr>
<td>Pair-wise Variational</td>
<td>0.083</td>
<td>0.502</td>
</tr>
<tr>
<td>Interpolated Smoothing</td>
<td>0.124</td>
<td>0.601</td>
</tr>
</tbody>
</table>
Experimental Results: Scientific Article Retrieval

• Used the relevance judgments produced for the 2015 TREC-CDS topics

• Same measures reported by NIST:
  – Inferred average precision (iAP)
  – Inferred normalized discounted cumulative gain (iNDCG)
  – R-precision
  – Precision at 10 (P@10)
## Experimental Results: Scientific Article Retrieval

<table>
<thead>
<tr>
<th>Q/A-CDS</th>
<th>iAP</th>
<th>iNDCG</th>
<th>R-Prec</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>*Interpolated</td>
<td>.147</td>
<td>.434</td>
<td>.344</td>
<td>.722</td>
</tr>
<tr>
<td>Bethe Approximation</td>
<td>.140</td>
<td>.432</td>
<td>.336</td>
<td>.701</td>
</tr>
<tr>
<td>Pair-wise</td>
<td>.128</td>
<td>.382</td>
<td>.330</td>
<td>.610</td>
</tr>
</tbody>
</table>

### Q/A-CDS

<table>
<thead>
<tr>
<th>Method</th>
<th>iAP</th>
<th>iNDCG</th>
<th>R-Prec</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>* BM25</td>
<td>.042</td>
<td>.204</td>
<td>.163</td>
<td>.387</td>
</tr>
<tr>
<td>TF-IDF</td>
<td>.041</td>
<td>.197</td>
<td>.169</td>
<td>.350</td>
</tr>
<tr>
<td>LMJM</td>
<td>0.040</td>
<td>.193</td>
<td>.151</td>
<td>.357</td>
</tr>
<tr>
<td>LMDir</td>
<td>0.043</td>
<td>.203</td>
<td>.170</td>
<td>.360</td>
</tr>
<tr>
<td>DFR</td>
<td>0.039</td>
<td>.197</td>
<td>.167</td>
<td>.333</td>
</tr>
</tbody>
</table>

### SotA

<table>
<thead>
<tr>
<th>Task</th>
<th>iAP</th>
<th>iNDCG</th>
<th>R-Prec</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task A Manual</td>
<td>–</td>
<td>.311</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Task B Manual</td>
<td>–</td>
<td>.381</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Task A Automatic</td>
<td>–</td>
<td>.294</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Task B Automatic</td>
<td>–</td>
<td>.382</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Presentation Outline

1. Introduction
2. System Architecture
3. Inferring the Answers
4. Experimental Results
5. Conclusions
Conclusions

• Both **Bethe Approximation** and **Interpolated Smoothing** produced relevant answers

• Best answers were obtained with the **extended medical sketch** \( EZ(t, l) \)

• Question Answering (Q/A) can be applied to CDS with very promising results

• Combining Q/A with IR can produce better results than using IR alone
Future Work

• MLE estimation produces the most commonly documented answers for a medical topic
• However, the medical topics used in TREC focused on difficult medical cases
• Incorporate “novelty” into the CPTG
  – We would like to see unusual answers as well as the typical answer
• Incorporate additional context into the CPTG
  – Scientific articles report on counter-indicated treatments, which are not negated in the text.
Questions?
Recognizing Medical Concepts

• Followed the framework of (Roberts and Harabagiu, 2010) to automatically recognize medical concepts.
• Relied on the 72,846 annotations provided in the 2010 i2b2 shared task.

3-stage classification:
1. A conditional random field (CRF) determines medical concept boundaries
2. A support vector machine (SVM) classifies each medical concept as a problem, treatment, or test.
3. Project each medical project onto UMLS and use the semantic type to distinguish between diagnoses, or signs/symptoms (e.g. SYMPTOM OR SIGN; FINDING)
Recognizing Medical Assertions

• Also followed the framework of (Roberts and Harabagiu, 2010) as well as the extension reported in (Goodwin and Harabagiu, 2012).
• Uses a support vector machine (SVM) to assign possible assertion values.
• 2010 i2b2 assertions:
  – PRESENT, ABSENT, POSSIBLE, HYPOTHETICAL, CONDITIONAL, or ASSOCIATED-WITH-SOMEONE-ELSE
  – Only apply to medical problems (diagnoses or signs/symptoms)
• We considered additional assertions:
  – For TREATMENTS, we included: PRESCRIBED, ONGOING, SUGGESTED
  – For TESTS, we included: CONDUCTED.
# Experimental Results: Answers II

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