CliniQA: Highly Reliable Clinical Question Answering System

Ni Yuan, Zhu Hui Jia, Cai Peng, Cao Feng, Zhang Lei, Qiu Zhao

INTRODUCING FAR REACHING RESEARCH
From IT perspective, one of the bottleneck of EBM is the difficulty for physician’s to effectively obtain relevant medical research results for a specific clinical question in hand.

Goal: build a system for physicians that can address clinical questions with highly reliable answers from latest state-of-the-art medical research results. Example:

Q: Is there any evidence for the use of Digoxin to reduce the mortality in patients with heart failure?

A: An RCT from The New England Journal of Medicine concluded that “Digoxin did not reduce overall mortality, but it reduced the rate of hospitalization both overall and for worsening heart failure”.

<table>
<thead>
<tr>
<th>Trial type</th>
<th>randomized, double-blinded, controlled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Followup</td>
<td>37 months</td>
</tr>
<tr>
<td>Problem</td>
<td>Patients with heart failure</td>
</tr>
<tr>
<td>Intervention</td>
<td>Digoxin</td>
</tr>
<tr>
<td>Comparison</td>
<td>Placebo</td>
</tr>
</tbody>
</table>

Table 2. Deaths According to Study Group and Cause.

<table>
<thead>
<tr>
<th>CAUSE OF DEATH</th>
<th>DIGOXIN (N=3397)</th>
<th>PLACERBO (N=3403)</th>
<th>ABSOLUTE DIFFERENCE</th>
<th>RISK RATIO</th>
<th>95% CI</th>
<th>P VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1181 (34.8)</td>
<td>1194 (35.1)</td>
<td>-0.4</td>
<td>0.99</td>
<td>0.91-0.97</td>
<td>0.80</td>
</tr>
<tr>
<td>Cardiovascular</td>
<td>1016 (29.9)</td>
<td>1004 (29.5)</td>
<td>0.4</td>
<td>1.01</td>
<td>0.93-1.10</td>
<td>0.76</td>
</tr>
<tr>
<td>Worsening heart failure</td>
<td>394 (11.6)</td>
<td>449 (12.2)</td>
<td>-1.6</td>
<td>0.88</td>
<td>0.77-1.01</td>
<td>0.06</td>
</tr>
<tr>
<td>Other cardiac</td>
<td>58 (1.7)</td>
<td>44 (1.3)</td>
<td>1.9</td>
<td>1.14</td>
<td>1.01-1.30</td>
<td>0.30</td>
</tr>
<tr>
<td>Other vascular</td>
<td>59 (1.5)</td>
<td>45 (1.3)</td>
<td>0.1</td>
<td>1.11</td>
<td>0.74-1.66</td>
<td>0.22</td>
</tr>
<tr>
<td>Unknown</td>
<td>64 (1.9)</td>
<td>66 (1.9)</td>
<td>-0.1</td>
<td>0.97</td>
<td>0.69-1.37</td>
<td>0.37</td>
</tr>
<tr>
<td>Noncardiac and nonvascular</td>
<td>165 (4.9)</td>
<td>190 (5.6)</td>
<td>-0.7</td>
<td>0.87</td>
<td>0.71-1.07</td>
<td>0.18</td>
</tr>
<tr>
<td>Method of assessing ejection fraction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radionuclide ventriculography</td>
<td>65.0</td>
<td>64.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-dimensional echocardiography</td>
<td>29.5</td>
<td>30.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrast angiography</td>
<td>5.5</td>
<td>5.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cardiothoracic ratio &gt;0.55</td>
<td>34.6</td>
<td>34.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NYHA class</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>13.7</td>
<td>13.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>53.3</td>
<td>54.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>30.7</td>
<td>30.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>2.2</td>
<td>1.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
System Architecture – Reuse DeepQA’s (Watson) Architecture

- Question Analysis
- Evidence Retrieval
- Evidence Scoring
- Final Ranking & Answer Generation
- Answer Sources with Annotations
- Result Display
- Answer & Confidence & Evidence object

Learned Models help combine and weigh the Evidence
A set of templates for therapy questions

The templates are defined in terms of the PICO framework, i.e. Problem (P), Intervention (I), Comparison (C), Outcome (O)

- `<I?, P>` asks for the intervention for a given problem, e.g. “how should I treat polymenorrhea in a 14-year-old girl?”
- `?<I, P>` asks for the evidences or the effectiveness of using the Intervention on the Problem, e.g. “what is the evidence for using Metformin in people with type 1 diabetes who are obese and poorly controlled?”
- `<I?>` asks for the usage of some intervention, e.g. “Is melatonin good for anything? I do not know anything about melatonin. I need to know the dose”

- Use MetaMap to identify the medical concepts from the questions
- Map the concepts into PICO categories in terms of the types
- Use the SemRap to detect whether a TREAT relationship exists
Semantic analysis to generate evidence object

- Identify P, I, C, O elements
  - make use of the linguistic and structure characteristics of the text, for Intervention
  - if the title satisfies the pattern “[XXX] of [YYY] on/for/in [ZZZ]”
  - if there exists a sentence in the objective part of the abstract section, and the sentence satisfies the pattern “the evaluation/effectiveness/impact/effect(s) of [YYY]”
  - the most frequently appeared medical concept of semantic type [TREATMENT&DRUG] is also considered as the candidates

- Identify randomization & followup of the clinical trials
- Identify the useful tables

Indexing & Searching

- Use Lucene to build the indexes
- Given the search query from the question analyzer, if the question has the requirements on the specific field, the retrieval will be conducted at the corresponding field
Component Technologies – Candidate Scorers

- **Bleu Scorer** is used to measure the similarity between the question and the matched passages.
- **Match Coverage Scorers** is to measure the percentage of matched concepts in the question.
- **Patient Gender Scorers** is to measure the similarity between the patient genders mentioned in the questions and the patient gender mentioned in the evidences.
- **Patient Age Scorers** is to measure the similarity between the patient ages mentioned in the questions and the patient ages mentioned in the evidences.
- **Disorder Similarity Scorers** is to measure the similarity between the disorders mentioned in the questions and the disorders mentioned in the evidences.
- **Drug Similarity Scorers** is to measure the similarity between the drugs mentioned in the questions and the drugs mentioned in the evidences.
- **1-gram, 2-gram, 3-gram**
- **Search rank**
Use machine learning to integrate all scores to get the final ranking

Extract the Conclusion part of the evidence as the answers
is Etoricoxib effect for osteoarthritis

The Answer:

Title: Etoricoxib in the treatment of osteoarthritis over 52-weeks: a double-blind, active-comparator controlled trial (NCT00242498) (E411)

Conclusion: In this extension study, etoricoxib, at doses ranging from 30 to 90 mg, demonstrated a maintenance of significant clinical efficacy in patients with OA through 52 weeks of treatment. Etoricoxib displayed clinical efficacy similar to diclofenac 150 mg and was generally well tolerated.

<table>
<thead>
<tr>
<th>Patients</th>
<th>Age</th>
<th>Gender</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40 years, 62 years</td>
<td>female</td>
<td>osteoarthritis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Interventions</th>
<th>Intervention</th>
<th>Trial</th>
<th>Follow-up</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Etoricoxib</td>
<td>877</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>The base study was 14 weeks in duration and consisted of 2 parts: in Part I (16 weeks), patients were allocated to once daily oral etoricoxib 5, 10, 30, 60, 90 mg or placebo. In Part II, (8 weeks), the placebo, etoricoxib 5 and 10 mg groups were reallocated to etoricoxib 30, 60, or 90 mg qd or diclofenac 50 mg t.i.d. Treatment was continued for consecutive 12 and 26 week extensions.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Aspect</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

IBM QL
Prototype Implementation – UIMA Framework

- The UIMA framework is used in the system
- The data pass between components are encapsulate as a CAS (Common Analysis Structure)
- Each component extends the CasAnnotator in UIMA
- Different components are arranged as a pipeline
UIMA Pipeline (1)

Evidence articles

XsgAnno → AgeAnno → GenderAnno → SnomedAnno → PICOAnno → TrialAnno → LuceneIndexer → ArticleCas Indexes

QuestionCas

XsgAnno → AgeAnno → GenderAnno → SnomedAnno → LuceneSearcher → BleuScorer → PatientGenderScorer → MatchCoverageScorer

DisorderSimilarityScorer → DrugSimilarityScorer → QuestionCas WithFeatures
TripAnswers dataset is leveraged
- 6265 questions with human generated answers

Example
- Question: Are oral B12 supplements in high doses equal effective as B12 intramuscular injections?
- Answer: In 2005 Cochrane published the systematic review ‘Oral vitamin B12 versus intramuscular vitamin B12 for vitamin B12 deficiency’ [1]. This concluded: “The evidence derived from these limited studies suggests that 2000 mcg doses of oral vitamin B12 daily and 1000 mcg doses initially daily and thereafter weekly and then monthly may be as effective as intramuscular administration in obtaining short term haematological and neurological responses in vitamin B12 deficient patients.” In the full-text of the review, they report: “Vitamin B12 replacement has been traditionally administered intramuscularly. However several case control and case series studies have since suggested equal efficacy and safety of the oral route (Chalmers 1958; Ross 1954; Spies 1949). The mechanism for this oral route is most probably that free vitamin B12 can be absorbed both passively (without binding to IF) as well as actively (following binding to IF) in the terminal ileum. Passive diffusion accounts for 1.2% of total absorption with a bioavailability unaffected in patients with pernicious anaemia or gastro-duodenal surgical resection (Berlin 1968; Berlin 1978). High doses of oral vitamin B12 (e.g. 1,000 micrograms daily) may be able to produce adequate absorption of vitamin B12 even in the presence of IF deficiency and therefore be an alternative to the intramuscular route in many patients.”
A simplified explanation of the Put-back experiment:

- TripAnswers Dataset = { <Q1, A1>, <Q2, A2>, … <Qn, An> }
- The Test Corpus = { A1, A2, … , An }
- The Test Question Set is a subset of {Q1, Q2, …, Qn}.
- For each test question Qi, CliniQA system searches the Test Corpus, and assign answers to the question. We simply check whether answer Ai is assigned in top positions to Qi.
Preliminary Results

- We randomly selected 1000 questions for training and 500 questions for test.
- The selected question set has the same distribution of different kinds of questions (i.e. drug, disorder, drug&disorder, other) as the original question set.

Metrics

  > Precision-recall curve
  
  - Sort the test questions based on their top answer’s confidence.
  - For each question at position i, suppose from question 1 to question i, the total number of correct answers are $c_i$, then we get a precision $c_i/i$ at recall $i/n$.
  - A curve is obtained for all precision – recall pairs.
Performance Result

- 0418: first base performance we obtained in April.
- 0705: with improved annotation and four new scorers, i.e. MatchCoverageScorer, DisorderSimilarityScorer, DrugSimilarityScorer, PersonGenderScorer
Thanks!

Q & A
What is the evidence for twice weekly cialis treatment to help with erectile dysfunction post prostatectomy?
GenerateArff:

CAS with CaFeatures → MapTo Weka Instances → Weka Instances Map → Feature Post Processing → Weka Instances → ArffSaver

Trainer:

ARFF data → ArffLoader → Weka Instances → Train Model (may have feature processing) → Models

Applier:

CAS with CaFeatures → MapTo Weka Instances → Weka Instances Map → Feature Post Processing → Weka Instances → Apply Model

Ranked answers with Confidences