Clinical Decision Support Track Overview

Matthew Simpson
Ellen Voorhees
William Hersh
Motivation for CDS Track

• Major emphasis on Health IT
  • goal to improve patient outcomes and reduce costs

• Clinical decision support systems
  • one piece of the target infrastructure
  • aim to anticipate physicians’ needs by linking health records to information needed for patient care
  • some of that info comes from biomedical literature

• Existing biomedical literature immense, and growth accelerating
  • difficult/impossible for clinicians to keep abreast
CDS Track Task

Given a case narrative, return biomedical articles that can be used to accomplish one of three generic clinical tasks:

- What is the diagnosis?
- What is the best treatment?
- What test should be run?

[Note: For the systems, this is an ad hoc document retrieval task, not a question answering task.]
CDS Track Task

Documents:

- open access subset of PubMed Central, a digital database of freely-available full-text biomedical literature
- track used subset as defined on Jan 21, 2014
- contains 733,138 articles in NXML
- images and other supplementary material available, though not included in basic release
CDS Track Task

- 30 topics
  - case narratives plus label designating which basic clinical task the topic pertains to
  - developed by physicians at NIH
  - 10 topics for each clinical task type
  - each topic statement includes both a “description” of the problem and a shorter, more focused “summary”
  - case narratives used as an “idealized” medical record since no collections of actual medical records available for use
<table>
<thead>
<tr>
<th>Topic Number</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Diagnosis</td>
<td>A 58-year-old nonsmoker white female with mild exertional dyspnea and occasional cough is found to have a left lung mass on chest x-ray. She is otherwise asymptomatic. A neurologic examination is unremarkable, but a CT scan of the head shows a solitary mass in the right frontal lobe.</td>
</tr>
<tr>
<td>13</td>
<td>Test</td>
<td>A 30-year-old generally healthy woman presents with shortness of breath that had started 2 hours before admission. She has had no health problems in the past besides 2 natural abortions. She had given birth to a healthy child 3 weeks before. On examination, she is apprehensive, tachypneic and tachycardic, her blood pressure is 110/70 and her oxygen saturation 92%. Otherwise, physical examination is unremarkable. Her chest x-ray and CBC are normal.</td>
</tr>
</tbody>
</table>

| Summary | 58-year-old female non-smoker with left lung mass on x-ray. Head CT shows a solitary right frontal lobe mass. |
| Summary | 30-year-old woman who is 3 weeks post-partum, presents with shortness of breath, tachypnea, and hypoxia. |
Runs

- Ranked list of up to 1000 docs per topic
- Standard two run types:
  - automatic: no human intervention from input of topic statement to output of ranked list
  - manual: everything else
- A given run must use the same topic type (summary vs. description) for all topics
- Max of 5 runs per participant
Participating Groups

26 groups participated in the track

Atigeo
Beijing U. of Posts and Telecommunications
BiTeM_SIBtex team, Geneva
The Chinese University of Hong Kong
CRP Henri Tudor
Dhirubhai Ambani Institute of Information & Communication Technology
East China Normal University
Georgetown University (2 groups)
Indian Institute of Technology, Varanasi
Institute of Medical Informatics, NCKU
JHU Human Language Technology CoE
Korea Institute of Science & Technology Information
LIMSI-CNRS
Medical Imaging Informatics, UCLA
Merck KGaA
Oregon Health & Science University
Philips
San Francisco State University
Seoul National University College of Medicine
University of Delaware
University of Michigan
Universidade Nova Lisboa
University of Texas at Dallas
Vienna University of Technology
York University
Relevance Judgments

- Judgments made by physicians
  - process overseen by OHSU
  - judge generally not topic author

- Judgment sets based on stratified samples
  - documents in top 20 ranks from all 102 runs, plus
  - 20% random sample of the set of docs retrieved between ranks 21–100 inclusive by some run
  - 37,949 topic-doc pairs to be judged
    (min: 908, max: 1669, mean: 1264.97 over topics)

- All docs in set judged on three-way scale
  - not relevant, possibly relevant, definitely relevant

- 8 topics fully double-judged
Per-Topic Prec(10) Scores

Topic

P(10) scores across all runs per topic

Text REtrieval Conference (TREC)
## Notable Topics

### Easiest (best median & best best infNDCG score)
- 4: 4-year-old boy with fever, conjunctivitis, strawberry tongue, desquamation of the fingers and toes [diagnosis]
- 9: soft, flesh-colored, pedunculated lesions on neck [diagnosis]

### Hardest (worst median & worst best infNDCG score)
- 23: heavy smoker with productive cough, shortness of breath, tachypnea, and oxygen requirement [treatment]
- 11: severe right arm pain and hypotension [test]

### Large differences between best & median infNDCG
- 5: shortness of breath 3 weeks after surgical mastectomy [diagnosis]
- 21: progressive arthralgias, fatigue, and butterfly-shaped facial rash [treatment]
Evaluation of Top Runs

Best run as measured by mean infNDCG(100) for top 8 groups

![Box plot showing the distribution of infNDCG values for the top 8 groups. The box plot compares different runs, with the manual run highlighted in blue.](image-url)
Evaluation of Top Runs

Best run as measured by mean Prec(10) for top 8 groups

Best Run By Mean P(10)
For many teams, a summary run is better than the corresponding description run, but best mean infNDCG run overall is a description-based run.
Dual-judged Topics

- 8 topics independently judged by two assessors
- Overlap of relevance sets on low side
  - but not outside of bounds seen in previous studies
  - lack of high-overlap single topics, but sample is small
- Anecdotal evidence confirms clinicians vary in their opinions of salient facts

<table>
<thead>
<tr>
<th>Topic</th>
<th>NN</th>
<th>NR</th>
<th>RR</th>
<th>RN</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1349</td>
<td>32</td>
<td>35</td>
<td>47</td>
<td>0.3070</td>
</tr>
<tr>
<td>5</td>
<td>1360</td>
<td>1</td>
<td>14</td>
<td>119</td>
<td>0.1045</td>
</tr>
<tr>
<td>12</td>
<td>838</td>
<td>17</td>
<td>114</td>
<td>508</td>
<td>0.1784</td>
</tr>
<tr>
<td>17</td>
<td>1040</td>
<td>53</td>
<td>13</td>
<td>6</td>
<td>0.1806</td>
</tr>
<tr>
<td>19</td>
<td>977</td>
<td>25</td>
<td>70</td>
<td>134</td>
<td>0.3057</td>
</tr>
<tr>
<td>25</td>
<td>1351</td>
<td>70</td>
<td>28</td>
<td>6</td>
<td>0.2692</td>
</tr>
<tr>
<td>27</td>
<td>437</td>
<td>17</td>
<td>296</td>
<td>158</td>
<td>0.6285</td>
</tr>
<tr>
<td>28</td>
<td>1070</td>
<td>10</td>
<td>35</td>
<td>17</td>
<td>0.5645</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.3173</td>
</tr>
</tbody>
</table>
Conclusion

- First year of CDS track
  - retrieval results suggest retrieval task is challenging, but doable

- User model
  - some debate over realism of the user task
    - in any case, technology developed will have wider applicability than track’s task
  - demonstration that clinical decisions at least as subjective as other relevance decisions