Clinical Corpus Annotation: Challenges and Strategies

Fei Xia and Meliha Yetisgen-Yildiz
University of Washington
Outline

• Motivation

• Related work - Three annotation schemata

• Our clinical NLP projects and corpora

• Challenges and strategies

• Conclusion
Clinical Natural Language Processing

- Electronic Medical Record (EMR) systems have become an integral part of health care services.
- Accessibility to the details of patient data available in EMR systems is critical to:
  - Improve health care process
  - Advance clinical research
- Most patient information is in free-text
Manual Annotation

• Roberts et al. summarized the reasons for the need of manual annotation as (Roberts et al 2009):
  – creating annotation scheme serves to focus and clarify the information requirements of the text processing task and the domain of interest
  – annotated data provides a gold standard to assess the performance of the text processing systems
  – annotated data serves as a resource for developing rule-based systems or creating statistical models by the application of machine learning approaches.
Three Annotation Schemata

• Traditional annotation
  – Gold standard is created by a small team
  – Example: most large-scale annotated corpora in the NLP field
  – Pros: high-quality, can handle complex tasks
  – Cons: long development cycle, expensive

• Crowd-sourced annotation
  – Gold standard is created by online labors
  – Example: many recent studies with Amazon’s Mechanical Turk
  – Pros: cheap, fast
  – Cons: simple tasks only, quality varies, need close monitoring

• Community annotation:
  – Gold standard is created by the community
  – Example: the i2b2 challenge
  – Pros: cheap, faster than traditional annotation
  – Cons: it works only in certain settings
Annotating Biomedical Data

• Biomedical literature
  – Information extraction tasks on biological events, entities, and their interactions (e.g., GENIA, PennBioIE, Yapex, GENETAG, …)

• Clinical text
  – Information extraction tasks for medical applications (e.g. ICD-9 coding, phenotype extraction, summarization …)
  – Not many publicly available corpora available due to privacy concerns (e.g., i2b2 challenge corpora on smoking history, comorbidities, medical entities, relations, assertions, …)
Our Clinical NLP Projects and Corpora

1. Critical recommendations in radiology reports

2. PNA and CPIS in chest X-ray reports

3. Pneumonia in the ICU reports
Project #1: Critical Recommendations in Radiology Report

- Clinical Collaborators: Dr. Martin Gunn (UW Medicine Department of Radiology), Dr. Tom Payne (UW Medicine ITS)

- Motivation:
  - Failure to communicate of abnormal test results are responsible for a large number of
    - medical errors
    - adverse events
    - liability claims
  - If a radiologist makes a potentially important observation, he/she may make further specific recommendations for follow-up imaging tests, or clinical follow-up in the report.

- Goal:
  - To automatically identify sentences that include recommendation information.
### Test: CT ABDOMEN and PELVIS

**Reason:** Prostate CA Surveillance

#### Incidental 6-mm left lung nodule.

*Follow-up chest CT is recommended in 6 months.*
Radiology Report Corpus

• 800 radiology reports
  – Reports represented a mixture of imaging modalities, including radiographs, CT scans, ultrasounds, MRIs

• Annotations are at the sentence level

• Two annotators went over each report and labeled the sentences with critical recommendation information
Project #2: PNA and CPIS in Chest X-ray Reports

• Clinical Collaborator: Dr. Heather Evans, UW Medicine Department of Surgery

• **Motivation:**
  – Early detection and treatment of VAP, one of the most common health care associated infections in critically ill patients is important
    • Short delays in appropriate antibiotic therapy are associated with higher mortality rates, longer term ventilation, and excessive hospitals costs

• **Goal:** Building an automated approaches to identify patients who are developing VAP based on structured data and free-text clinical reports
Chest X-ray Corpus

• We created a corpus of 1344 chest x-ray reports

• Annotations are at the **report level**

• Two annotators annotated each report with
  – Clinical Pulmonary Infection Score (CPIS)
    • 1 – no infiltrate
    • 2 – diffuse infiltrate
    • 3 – localized infiltrate
  – Pneumonia
    • 1 – no suspicion (negative for pneumonia)
    • 2 – suspicion of pneumonia
    • 3 – probable pneumonia (positive for pneumonia)

• The annotators also highlighted the text in the reports that support their decision.
Project #3: Pneumonia in the ICU reports

• Clinical Collaborator: Dr. Mark Wurfel, UW Medicine Division of Pulmonary and Critical Care Medicine

• **Motivation:**
  – Identification of complex critical phenotypes among critically ill patients is a major challenge
  – Current approach - Manual chart abstraction
    • Labor intensive
      – Requires examination of various report types: Admit notes, ICU daily progress notes, discharge summaries, …
    • Involves many subjective assessments

• **Goal:** To apply NLP and ML to accurately identify critical illness phenotypes and their interactions among ICU patients
Penumonia Corpus

• Annotations are at patient level
• A cohort of 426 ICU patients
  – 66 cases for positive pneumonia
  – 360 cases for negative pneumonia
• A total of 5,313 reports

<table>
<thead>
<tr>
<th>Report type</th>
<th>Reports</th>
<th>Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Admit note</td>
<td>481</td>
<td>280</td>
</tr>
<tr>
<td>ICU daily progress note</td>
<td>2,526</td>
<td>388</td>
</tr>
<tr>
<td>Acute care daily progress note</td>
<td>1,357</td>
<td>203</td>
</tr>
<tr>
<td>Interim summary</td>
<td>164</td>
<td>115</td>
</tr>
<tr>
<td>Transfer/transition note</td>
<td>243</td>
<td>175</td>
</tr>
<tr>
<td>Transfer summary</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Cardiology daily progress note</td>
<td>133</td>
<td>17</td>
</tr>
<tr>
<td>Discharge summary</td>
<td>391</td>
<td>350</td>
</tr>
</tbody>
</table>
## Summary of Three Corpora

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Report Type</th>
<th>Corpus Size</th>
<th>Annotation</th>
<th>Annotation Unit</th>
<th>Annotators</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>Radiology reports</td>
<td>800 reports</td>
<td>Critical recommendation</td>
<td>Sentence</td>
<td>1 radiologist, 1 internal medicine physician</td>
</tr>
<tr>
<td>C2</td>
<td>Chest x-ray reports</td>
<td>1344 reports</td>
<td>PNA and CPIS</td>
<td>Report</td>
<td>1 surgeon, 1 data analyst</td>
</tr>
<tr>
<td>C3</td>
<td>Eight ICU report types</td>
<td>5313 reports for 426 patients</td>
<td>PNA</td>
<td>Patient</td>
<td>1 research RN</td>
</tr>
</tbody>
</table>
Challenges

• Annotation tasks
  – rely on medical expertise
  – requires protection of patients’ privacy

• Traditional annotation schema doesn’t fit well
Traditional Annotation Schema

1. Define annotation task based on the need of the community (l, c)
2. Select data to be annotated (l)
3. Write a detailed set of annotation guidelines (d, e)
4. Create good annotation tools (l, t)
5. Find and train annotators (l)
6. Annotate text
   a) Annotate text based on the guidelines (a)
   b) Revise annotation guidelines if needed (d, e)
   c) Monitor inter-annotator agreement and re-train annotators (l)
   d) Modify annotation based on the revised guidelines (a)
   e) Once some data have been annotated, train some NLP systems to pre-process the data to speed up annotation (l, t)
7. Release the corpus to the community (l)
8. Use the corpus to build various systems (c)
9. Find additional funding to extend the corpus, repeat some of the previous steps (l)

l: leader, d: guideline designer, e: domain expert, a: annotator, t: technical support
Characteristics of Clinical Annotation

• Annotation by experts:
  – How much knowledge is required? And how soon the annotators can acquire such knowledge?
  – For the projects described, medical expertise is a must for both:
    • Design of annotation guidelines
    • Annotation process
# Characteristics of Pneumonia

<table>
<thead>
<tr>
<th>Causes</th>
<th>Clinical Signs and Symptoms</th>
<th>Risk Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- <em>H. influenza</em></td>
<td>Fever</td>
<td>- Age &gt; 65</td>
</tr>
<tr>
<td>- <em>Strep pneumonia</em></td>
<td>Sputum production</td>
<td>- Immunosuppression</td>
</tr>
<tr>
<td>- <em>Staph aureus</em></td>
<td>Cough</td>
<td>- Recent antibiotic use</td>
</tr>
<tr>
<td>- Legionella species</td>
<td>Shortness of breath</td>
<td>- Comorbid illnesses: HIV, Asthma, COPD, Renal Failure, CHF, Diabetes, Liver Disease, Cancer, Stroke</td>
</tr>
<tr>
<td>- Chlamydia species</td>
<td>Chest Pain</td>
<td></td>
</tr>
<tr>
<td>- <em>Pseudomonas aeruginosa</em></td>
<td>Abnormal white blood cell count</td>
<td></td>
</tr>
<tr>
<td>- *Viruses:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Influenza</td>
<td>Malaise, fatigue</td>
<td></td>
</tr>
<tr>
<td>- Parainfluenza</td>
<td>Muscle pains</td>
<td></td>
</tr>
<tr>
<td>- *Fungi:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Blastomycosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Coccidiomycosis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Histoplasmosis</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Characteristics of Clinical Annotation (Cont.)

• Impact of privacy considerations
  – IRB review can take a long time
    • Six weeks to two months at UW
  – The application needs to be prepared carefully
    • Time extension is not allowed at UW
  – It is often difficult to get approval to release the annotated corpus
    • Publicly available corpora: i2b2 NLP challenges and TREC Medical Records Track
Characteristics of Clinical Annotation (Cont.)

• Impact of legal considerations
  – Malpractice lawsuits
  – Hedging:
    • Defined as an evasive statement to avoid the risk of commitment
    • Presents a source of ambiguity from the perspective of annotation and leads to annotation disagreement
      – Example: *If clinically indicated pelvic ultrasound could be performed in 4 to 6 weeks to document resolution.*
Effects on the Annotation Process

• Roles of NLP researchers
  – Typical annotation project: NLP researchers often play a central role
  – Clinical NLP projects: Their role becomes limited due to lack of medical knowledge
Effects on the Annotation Process (Cont.)

• Guidelines
  – Physicians are not familiar with the common practice of annotation such as creating detailed annotation guidelines in advance and revising guidelines if necessary
Effects on the Annotation Process (Cont.)

• Finding and training annotators
  – Compared to annotators, physicians are more expensive and have very busy schedules

  – Training and re-training physicians is difficult since the disagreements could be due to different interpretations related to different medical domains
Process of Expert Annotations

1. Define an annotation task based on the clinical needs (physician)
2. Select data to be annotated (physician)
3. Get IRB approval (physician, nlp researcher)
4. Write annotation guidelines (physician, nlp researcher)
5. Create good annotation tools (nlp researcher)
6. Annotate text based on the guidelines and/or medical training (physician)
7. Use the corpus to build various systems (nlp researcher)
8. Test how well the systems meet the clinical needs (physician)
Strategies

1. Importance of annotation guidelines
   - Two rounds of annotation to test the importance of annotation guidelines
   - Round 1:
     • No annotation guideline
       – 1 radiologist and 1 internal medicine physician annotated C1
       – 1 surgeon and 1 data analyst annotated 100 documents from C2
   - Round 2:
     • Annotators + NLP researchers went over the disagreements and put together detailed annotation guidelines
       – C1: annotators went only over the disagreements
       – C2: annotators waited a few days and re-annotated the 100 documents
Annotation Guideline for CPIS (round 2)

1A: NO INFILTRATE
• The report includes information that neither diffuse nor localized infiltrate. The report could include edema or pleural effusion.
• If there are extra pleural mentions in the report, they are not related to PNA.

1B: DIFFUSE INFILTRATE OR ATELECTASIS
• Atelectasis is more important than localized process that is consistent with infection.
• Lobar collapse is consistent with atelectasis.
• Multiple areas of opacity could fall under 1B.
• If bi-basilar consolidation is present with bi-pleural effusion much more suggestive of atelectasis.

1C: LOCALIZED INFILTRATE
• If one opacity is specifically highlighted and PNA or infection also mentioned in text, than this is more important than 1A and 1B.
## Agreement Results – Radiology Corpus

<table>
<thead>
<tr>
<th>Round</th>
<th>A1</th>
<th>A2</th>
<th>Agreed</th>
<th>P/R/F</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>110</td>
<td>109</td>
<td>83</td>
<td>0.755/0.761/0.758</td>
<td>0.757</td>
</tr>
<tr>
<td>2</td>
<td>114</td>
<td>118</td>
<td>113</td>
<td>0.991/0.958/0.974</td>
<td>0.974</td>
</tr>
</tbody>
</table>

IAA for the **Radiology Corpus** (C1). The corpus has 800 documents and 18,748 sentences in total. The “A1” and “A2” columns show the number of critical recommendation sentences; the “Agreed” column shows the number of positive sentences marked by both annotators; P/R/F scores are precision, recall, and F-score for identifying positive sentences when A2’s annotation is treated as gold standard and A1’s annotation is treated as system output; “kappa” is the kappa coefficient.
## Agreement Results – Chest X-ray Corpus

Clinical Pulmonary Infection Score (CPIS)

<table>
<thead>
<tr>
<th>Round</th>
<th>A1</th>
<th>A2</th>
<th>Agreed</th>
<th>Acc</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13/59/28</td>
<td>15/74/11</td>
<td>12/52/6</td>
<td>70</td>
<td>0.415</td>
</tr>
<tr>
<td>2</td>
<td>13/72/15</td>
<td>16/72/12</td>
<td>13/68/10</td>
<td>91</td>
<td>0.797</td>
</tr>
</tbody>
</table>

Pneumonia (PNA)

<table>
<thead>
<tr>
<th>Round</th>
<th>A1</th>
<th>A2</th>
<th>Agreed</th>
<th>Acc</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13/59/28</td>
<td>15/74/11</td>
<td>12/52/6</td>
<td>45</td>
<td>0.085</td>
</tr>
<tr>
<td>2</td>
<td>67/19/15</td>
<td>67/32/1</td>
<td>13/68/10</td>
<td>85</td>
<td>0.697</td>
</tr>
</tbody>
</table>

IAA on **CPIS** and **PNA** labeling for the 100 double annotated reports in the chest X-ray corpus (C2). x/y/z in each cell of the “A1”, “A2”, and “Agreed” columns are the numbers of reports with labels 1a, 1b, and 1c, respectively; “Acc” is the percentage of reports that receive the same CPIS/PNA label from the two annotators; “kappa” is the kappa coefficient.
2. Providing additional information

– Choosing granularity of information is important for building successful systems

  • C3 includes 426 yes/no answers for pneumonia at the patient level

  • Information on the cues the annotators used to determine pneumonia cases is missing
Strategies (Cont.)

3. Time commitment from physicians
   - Physicians are very interested in clinical NLP projects
   - However, they underestimate the time required to build a high quality dataset
   - NLP researchers should explain the annotation process to physicians and together they need to decide an annotation plan
Strategies (Cont.)

4. Early involvement of NLP researchers

– Although NLP researchers’ involvement in the annotation process is less than in the traditional approach, they should get involved in the process as early as possible
  • Explaining the annotation process to physicians
  • Structuring the annotation guidelines with physicians
  • Calculating the inter-rater agreement levels
Conclusion

• Clinical NLP projects require expert annotation

• Without detailed guidelines/discussions the annotation agreement among experts is low

• NLP researchers play a minor role in providing content in guideline designs and annotation due to their lack of knowledge

• NLP researchers’ early involvement is important for the success of annotation
UW-BioNLP Group

- [http://depts.washington.edu/bionlp/index.html](http://depts.washington.edu/bionlp/index.html)
- Clinical Natural Language Processing Research
- Group Members:
  - Faculty:
    - Meliha Yetisgen-Yildiz (Biomedical Health Informatics)
    - Fei Xia (Department of Linguistics)
    - Lucy Vanderwende (Microsoft Research)
  - Post-doctoral Researcher:
    - Cosmin Adi Bejan (Biomedical Health Informatics)
  - Graduate Students:
    - Michael Tepper (PhD Student - Department of Linguistics)
  - Clinical Collaborators:
    - Thomas Payne, UW Medicine ITS
    - Martin Gunn, Radiology
    - Mark Wurfel, Pulmonary and Critical Care Medicine
    - Gail Rona, Pulmonary and Critical Care Medicine
    - Heather Evans, Surgery
Questions