Clinical Corpus Annotation: Challenges and Strategies

Fei Xia[1,2] and Meliha Yetisgen-Yildiz[2,1]
[1]Department of Linguistics, [2]Biomedical and Health Informatics
University of Washington, Seattle, WA 98195, USA
E-mail: {fxia,melihay}@uw.edu

Abstract

Annotation is an important task for Natural Language Processing (NLP), and the traditional annotation schema, including writing detailed guidelines and training annotators, has proved to work well in many previous annotation projects. However, making medical judgment on clinical data requires medical expertise and annotation can only be done by experts. Recently, we created three corpora for our clinical NLP studies: one marks critical recommendations in radiology reports, and the other two indicate whether a patient has pneumonia based on chest X-ray reports or ICU reports. All the annotations were done by medical experts. In this paper, we discuss various challenges we have encountered when dealing with expert annotation, and lay out some lessons we have learned from the annotation tasks. Our experiments show that medical training alone is not sufficient for achieving high inter-annotator agreement, and NLP researchers should get involved in the annotation process as early as possible despite their lack of medical training.

Keywords: Clinical corpus annotation, Annotation schemata, Annotation guidelines

1. Introduction

Over the last decade, Electronic Medical Record (EMR) systems have become increasingly integral to the provision of health care services. Accessibility to the details of patient data available in EMR systems is critical to improve the health care process and advance clinical research. However, most patient information that describes patient state, diagnostic procedures, and disease progress is represented in free-text form. Several studies demonstrated the value of Natural Language Processing (NLP) in extracting the knowledge from clinical records for a variety of health care applications including decision support tools, quality improvement initiatives, and automated encoding for clinical research (Chapman and Cohen, 2009).

Although the premise of NLP is to develop automated approaches to process free-text data available in medical records, building those approaches requires a substantial amount of manual analysis and annotation of data. Roberts et al. (2009) summarized the reasons for the need of manual annotation as: (1) creating annotation scheme serves to focus and clarify the information requirements of the text processing task and the domain of interest, (2) annotated data provides a gold standard to assess the performance of the text processing systems, and (3) annotated data serves as a resource for developing rule-based systems or creating statistical models by the application of machine learning approaches. Therefore, producing high quality annotations is essential to building successful text processing systems.

The traditional annotation schema in the NLP field includes detailed annotation guidelines, well-trained annotators, double annotation, and adjudication. However, when the annotated data are clinical records and the annotation requires substantial medical expertise, we face new challenges while trying to follow the traditional annotation schema. In this paper, we discuss three annotation tasks that use clinical data, describe various challenges we encounter, and lay out some lessons we have learned from the tasks.

2. Related Work

In this section, we discuss common annotation schemata and related work on annotating biomedical data.

2.1 Three Annotation Schemata

Annotation is important for NLP research. Traditionally, an annotated corpus is created by a team consisting of guideline designers, annotators, language or domain experts, and technical support staff. Detailed annotation guidelines are created before annotation starts and they are revised during annotation if necessary. The annotators in this team are often trained on the annotation task for a long period of time. We refer to this approach as *traditional annotation* schema. Almost all the large-scale annotated corpora used in the NLP community were created this way, including the Prague Dependency Treebank (Hajic, 1999), the English/Chinese/Arabic Penn Treebank (Marcus et al., 1993; Xia et al., 2000; Maamouri and Bies, 2004), the English PropBank (Palmer et al., 2005), and the Penn Discourse Treebank (Miltsakaki et al., 2004).

One issue with the traditional annotation scheme is the high cost of training and maintaining annotators. Recently, two other annotation schemata are proposed to address this issue. The first one, *crowd-sourced annotation*, takes advantage of online labor markets such as Amazon’s Mechanical Turk (AMT). Because the cost of labors from such markets is much lower than that of well-trained annotators, corpus developers can afford to have multiple annotations on the same data and use majority voting to choose gold standard. This schema can produce good results with low cost when the task is relatively simple and does not require much domain knowledge (Snow et al., 2008). AMT has been applied to the biomedical domain successfully for named-entity extraction tasks from clinical trial descriptions (Yetisgen-Yildiz et al., 2010). The second schema, *community annotation*, gathers annotation from a research community; one example is the evaluation corpus used in the 2009 i2b2 medication challenge, which was created by the i2b2 organizers (who created annotation guidelines and some initial annotations) and the participating teams. This schema can produce good annotation fast and with low cost, if the annotation is careful coordinated and receives strong support from the community (Uzuner et al., 2010).

Corpus	Report Type	Corpus size	Annotation	Annotation Unit	Annotators
C1	Radiology reports	800 reports	critical recommendation	sentence	one radiologist, one internal medicine physician
C2	Chest x-ray reports	1344 reports	PNA and CPIS	report	one general surgeon, one data analyst
C3	Eight ICU report types	5313 reports for 426 patients	PNA	patient	one research study nurse

Table 1. The three corpora in our study. “PNA” stands for “pneumonia”, “CPIS” stands for “Critical Pulmonary Infection Score”, and “ICU” stands for “intensive care unit”.

2.2 Annotating Biomedical Data

Data in the biomedical domain can be divided into the two types. The first type is the collection of research articles in the biomedical literature such as Medline. The second type is clinical patient data such as radiology reports. For the first type, there are various corpora generated from the research articles available in Medline for information extraction tasks on biological events, entities and their interactions. Some well-known, publicly available biomedical corpora include GENIA corpus (Kim et al., 2003), PennBioIE corpus (Kulick et al., 2004), Yapex corpus (Franzen et al., 2002), and GENETAG (Tanabe et al., 2005).

For clinical data, the number of publicly available annotated corpora is quite limited due to concerns regarding patient privacy as well as concerns about revealing unfavourable institutional practices (Chapman et al., 2011). The i2b2 NLP challenges contribute to the clinical NLP research by releasing corpora composed of de-identified clinical records annotated for various different information extraction tasks including smoking history extraction (Uzuner et al., 2008), comorbidity extraction (Uzuner, 2009), named-entity extraction (medication, treatment, test, medical condition) (Uzuner et al., 2010; Uzuner et al., 2011), assertion and relation extraction (Uzuner et al., 2011).

There are other studies on annotating clinical data that are not publicly available. While some of those clinical corpora are about traditional NLP annotations such as POS tagging (Pakhomov et al., 2006) and anaphoric relations (Savova, 2011), other corpora require annotators to be medical experts. These annotation tasks are more domain specific, focusing on the annotation of medical knowledge in clinical text. One example is the Clinical E-Science Framework (CLEF) corpus (Roberts et al., 2007; Roberts et al., 2009). The purpose of the CLEF project is to build a framework for the capture, integration and presentation of clinical information to be used in clinical research, evidence-based health care, and genotype-phenotype mapping. The corpus includes various types of clinical records annotated for named entities and their relations, modifiers, and co-references. Because of the nature of the clinical research, most of the corpora generated in this domain are very specific to a disease or a disease type. For example, Fiszman et al. (2000) annotated chest x-ray reports for automatic identification of acute bacterial pneumonia; South et al. (2009) manually annotated clinical records to identify phenotypic information for inflammatory bowel disease. Fiszman et al.’s annotation was at the report level, whereas South et al.’s annotation was at the phrase level. These three corpora require significant medical knowledge, and the corpora we build in our research projects fall into this category.

3. Our Projects and Corpora

In this section, we discuss three projects that we are currently working on. For each project, we created a corpus to train and evaluate our NLP systems. All three projects deal with patient medical reports, and the corpora were annotated by physicians. The retrospective review of the reports in the corpora was approved by the Human Subjects Committee of Institutional Review Board (IRB) at our institute, who waived the need for informed consent. Table 1 provides a summary of the corpora. In the rest of the section, we will provide a background of the projects, a description of the corpora, and some preliminary results of our NLP systems.

3.1 Critical Recommendations in Radiology Reports

Radiology reports include the descriptions of relevant disease processes found by radiologists on imaging studies, such as radiographs and computed tomography (CT) scans. If a radiologist makes a potentially important observation when examining an imaging study, he/she may include in his/her report further specific recommendations for follow-up imaging tests, or clinical follow-up. These recommendations are made when the radiologist considers the finding to be clinically significant and unexpected, and believes that it is important for the referring physician to consider further investigation, management, or follow-up of the finding in order to avoid an adverse outcome. The American College of Radiology (ACR) recommends that radiologists supplement their written report with “non-routine” means of communication with the referring physician (usually verbal) to ensure adequate receipt of the critical information in a timely manner[1]. Despite the imperative of good communication to avoid medical errors, it does not always occur. Inadequate communication of critical results is the cause of the majority of malpractice cases involving radiologists in the USA (Towbin et al., 2011). The Joint Commission reported that up to 70% of sentinel medical errors were caused by communication errors (Lucey and Kushner, 2010).

[1] ACR practice guideline for communication of diagnostic imaging findings. Available at: http://www.acr.org/SecondaryMainMenuCategories/quality_safety/guidelines/dx/comm_diag_rad.aspx

The goal of our first project is to build an NLP system that automatically identifies critical recommendations in radiology reports so that these recommendations will be highlighted to reduce the chance that they are overlooked by the referring physicians. We defined *critical recommendation* as a statement made by the radiologist in a radiology report to advise the referring clinician to further evaluate an imaging finding by either other tests or further imaging. An example sentence annotated as critical recommendation from our corpus is "*Recommend non-emergent pelvic ultrasound for further evaluation to exclude cystic ovarian neoplasm*."

In order to train and evaluate our system, we created a corpus of radiology reports composed of 800 de-identified radiology reports extracted from Harborview Medical Center radiology information system. Two annotators, one radiologist and one clinician, went through each of the 800 reports and marked the sentences that contained critical recommendations. Out of 18,748 sentences in the reports, the radiologist annotated 118 sentences and the clinician annotated 114 sentences as recommendation. They agreed on 113 of the sentences annotated as recommendation.

Using the corpus, we built a statistical text processing system to classify each sentence in radiology reports as either containing or not containing critical recommendation. The system achieved 95.60% precision, 79.82% recall, and 87% F-score (5-fold cross validation) in identifying recommendation sentences. More detail of the system design and evaluation was reported in (Yetisgen-Yildiz et al., 2011a).

3.2 PNA and CPIS in Chest X-ray Reports

Early detection and treatment of ventilator associated pneumonia (VAP), the most common healthcare associated infections in critically ill patients, is important; even short-term delays in appropriate antibiotic therapy are associated with higher mortality rates, longer-term mechanical ventilation, and excessive hospital costs. Traumatic injury places patients at particular risk for VAP, and efforts to perform accurate risk assessment and diagnostic confirmation should be focused in this population. Interpretation of meaningful information from the EMR at the bedside is complicated by high data volume, lack of integrated data displays and text-based clinical reports that may be reviewed only by manual search. This cumbersome data management strategy obscures the subtle signs of early infection.

The goal of our second project is to build NLP systems to identify patients who are developing critical illnesses in a manner timely enough for early treatment. As a first step, we have built a system that determines whether a patient has pneumonia based on the narrative text of the patient's chest X-ray reports.

To train and evaluate the system, we created a corpus of 1344 chest X-ray reports from our institution. Two annotators, one is a general surgeon and the other is a data analyst in a surgery department, read each report and determine whether the patient has pneumonia (PNA) and also what the clinical pulmonary infection score (CPIS) is for the patient. The CPIS is used to assist in the clinical diagnosis of VAP by predicting which patients will benefit from obtaining pulmonary cultures. The use of the CPIS is shown to result in fewer missed VAP episodes and can also prevent unnecessary antibiotic administration due to treatment of colonized patients.[2] There are three possible labels for CPIS: (1a) no infiltrate, (1b) diffuse infiltrate or atelectasis, and (1c) localized infiltrate. There are also three possible labels for PNA: (2a) no suspicion (negative class), (2b) suspicion of PNA, and (2c) probable PNA (positive class). The difference between the labels (2b) and (2c) is the certainty level on PNA. If there is enough evidence in a given report that indicates PNA, the report is labeled with (2c). If the evidence in the report is not enough to label it with (2c) but also not enough to rule out the possibility of PNA (2a), then it is labeled with (2b).

We used this corpus to train two classifiers, one for CPIS and the other for PNA). We did 5-fold cross validation. The accuracy of the CPIS classifier was 85.86%. The accuracy of the PNA classifier was 78.2% for the 3-way distinction, and the performance improved to 85.19% for the 2-way distinction when the two codes indicating suspicion of pneumonia, (2b) and (2c), were collapsed into a single class.

3.3 Pneumonia in the ICU Reports

With the introduction of comprehensive EMRs, all aspects of intensive care unit (ICU) care are now captured in both structured and free-text format. The existence of such data provides an opportunity to identify critical illness phenotypes and facilitate clinical and translational studies of large cohorts of critically ill patients, a task that would not be feasible using traditional screening/manual chart abstraction methods.

The goal of our third project is to build automated tools to identify critical illness phenotypes such as pneumonia (PNA) and model their progression based on the ICU reports. PNA can be classified further based on the context in which it occurs. Community acquired pneumonia (CAP) refers to pneumonia that occurs outside of the hospital setting; whereas hospital acquired pneumonia (HAP) refers to pneumonia which occurs after admission to the hospital. VAP is a special case of HAP, where the infection can be linked to the use of the ventilation machine.

Physician daily notes are a potentially rich source of clinical information indicating the presence of phenotypes such as pneumonia. In contrast to the narrow scope of information provided by radiology reports (e.g., chest X-ray reports), physician daily notes include text detailing patient narrative, physiologic, imaging, and laboratory data, and, finally, the physician's interpretation of these data. We hypothesized that by using physician notes such as admit notes, ICU progress notes, and discharge summaries, automated approaches that incorporate NLP and machine learning can accurately identify pneumonia in ICU settings.

To train and evaluate our PNA detection system, we created a corpus composed of ICU reports for 426 patients. An annotator with 6 years of experience as a research study nurse manually classified a patient as "positive" if the patient had pneumonia within the first 48 hours of ICU admission and as "negative" if the patient did not have pneumonia or the pneumonia was detected after the first 48 hours of ICU admission (66 cases positive for pneumonia and 360 cases negative for

[2] http://www.surgicalcriticalcare.net/Resources/CPIS.php

pneumonia). The annotation was per-patient. Because subjects in this dataset were admitted to the ICU from the emergency department as well as from other hospitals, cases of pneumonia included both CAP and HAP. Table 2 provides a summary of the characteristics of pneumonia.

CAUSES	
Bacteria: - *H. influenza* - *Strep pneumonia* - *Staph aureus* - Legionella species - Chlamydia species - *Pseudomonas aeruginosa*	Viruses: - Influenza - Parainfluenza Fungi: - Blastomycosis - Coccidiomycosis - Histoplasmosis
CLINICAL SIGNS AND SYMPTOMS	
Fever	Sputum production
Cough	Shortness of breath
Chest Pain	Malaise, fatigue
Abnormal white blood cell count	Muscle pains
RISK FACTORS	
Age > 65	
Immunosupression	
Recent antibiotic use	
Comorbid illnesses: HIV, Asthma, COPD, Renal Failure, CHF, Diabetes, Liver Disease, Cancer, Stroke	

Table 2. Characteristics of Pneumonia

Our dataset includes a total of 5313 reports from eight report types (admit note, ICU daily progress note, acute care daily progress note, interim summary, transfer/transition note, transfer summary, cardiology daily progress note, and discharge summary) for 426 patients. The total number of reports per patient ranged widely (median=8, interquartile range = 5-13, minimum =1, maximum=198). This is due to the high variability in the length of ICU stay. The distribution among the eight different report types is presented in Table 3. The first column of the table gives the number of reports for each report type and the second column gives the number of distinct patients who had the report type in the dataset.

REPORT TYPE	REPORT COUNT	PATIENT COUNT
ADMIT NOTES	481	280
ICU DAILY PROGRESS NOTE	2526	388
ACUTE CARE DAILY PROGRESS NOTE	1357	203
INTERIM SUMMARY	164	115
TRANSFER/TRANSITION NOTE	243	175
TRANSFER SUMMARY	18	18
CARDIOLOGY DAILY PROGRESS NOTE	133	17
DISCHARGE SUMMARY	391	350

Table 3. Statistics of the ICU corpus. Report Count: The number of reports with that report type; Patient Count: The number of distinct patients who had that report type.

In (Yetisgen-Yildiz, 2011b), we presented the preliminary results of the statistical system we built to identify PNA trained with this corpus. With 5-fold cross validation, our classifier achieved 58.3% precision, 42.4% recall, and 49.1% F1 for identifying patients with PNA. The classification accuracy was 86.4% and the specificity was 94.4%.[3]

4. Challenges

Given the nature of our annotation tasks, which relies on the medical expertise of annotators and requires protection of patients' privacy, the crowd-sourced annotation or community annotation schemata would not be applicable. Ideally, we would want to follow the traditional annotation schema, which has been proved to work well in numerous projects; however, we encounter several challenges due to some characteristics of our annotation tasks, and as a result, we have to make some changes to the traditional annotation schema.

4.1 Traditional Annotation Schema

In the traditional annotation schema, the annotation is done by a team consisting of the following members: project leader (l), guideline designers (d), linguistics / domain experts (e), annotators (a), and technical support (t). In addition, the team will ask its large research community (c) for suggestions, feedback and support.

Below is a common procedure for the traditional annotation schema, and the people who are in charge of each step are shown in parentheses:

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)

4.2 Characteristics of Clinical Annotation

Compared to most annotation projects in the general domain or the biomedical domain, our projects differ in several ways.

4.2.1 Annotation by Experts

For any annotation task on a non-general domain, having domain knowledge is helpful for the annotation team. The question is how much knowledge is required and how soon an annotator can acquire such knowledge. In our projects, medical expertise is a must for both design of the annotation guidelines and annotation itself, and it cannot be acquired quickly. As a result, we have to heavily rely on medical experts. We call this kind of

[3] Specificity is the negative (non-PNA) predictive value.

annotation "expert annotation".

For instance, in the ICU corpus (C3), the annotator needs to go over all the ICU reports of a patient in order to determine whether the patient has pneumonia within 48 hours of admission to ICU. Very often, the ICU reports would not explicitly say whether or not the patient has pneumonia. The annotator, a research study nurse with six-year experience, has to use her medical expertise to determine whether the patient has any of the characteristics of the disease (see Table 2) and whether the identified characteristics are sufficient to make the call. For instance, when she sees the text "WBC: 15000 mcl" in a report, she knows that "WBC" stands for "white blood cell", "mcl" stands for "microliter", and the normal range of WBC count is 4,500-10,000 per microliter. So she knows that the text span indicates that the patient has "abnormal white blood cell count", a symptom under "Clinical Signs and Symptoms". Once she has found all the relevant cues in the text, she needs to then decide whether they are sufficient for her to label the patent as "having PNA". All this domain knowledge cannot be acquired by a layman in a short period of time (say within a few months). Similarly, annotation guidelines such as the one in Table 4 for the chest X-ray corpus (C2) can be created and understood only by medical experts trained in a particular field.

4.2.2 Impact of Privacy Consideration

When annotating clinical data, privacy is an important concern. In addition to the requirement of getting IRB approval in advance, there are other ramifications; two examples are given here:

- The IRB review process can take a long time, and no one can work on the data before the IRB is approved. This leads to less flexibility in selecting the data set and choosing annotators. For instance, in the ICU project (C3), after the IRB approval, we got access to the records of the 426 patients listed on the IRB form, and we then realized that some patients missed important reports such as discharge summaries. But at that time, it was already too late to request records for additional patients, because that would require a new IRB approval, which could take additional time depending on the institution. Similarly, an annotator cannot work on a project unless the request of adding him/her to the project has been approved by the IRB.
- It is often very difficult for the annotation team to get approval to release the corpus to the research community. In the United States, HIPAA[4] provides guidelines for protecting patient information. HIPAA considers the data to be de-identified if the data is cleaned of seventeen categories of possible identifiers including personal health information (PHI) and any other information that may make it possible to identify the individual. Therefore, even if the corpus can be released, the de-identification process would make the corpus less useful for research purpose. If a corpus cannot be released, it becomes impossible for the community to benefit from the corpus and for the annotation team to get feedback from the community.

4.2.3 Impact of Legal Considerations

One characteristic of clinical domain is the concern about malpractice lawsuits. Let us use the radiology report corpus (C1) as an example. Poor communication has been found to be a causative factor in up to 80% of malpractice lawsuits involving radiologists (Levinson, 1994). In those lawsuits, the radiology report is often treated as an important medico-legal document. Given the legal aspect of the reports, it is common for a radiologist to use "hedging" in their reports (Wallis and McCoubrie, 2011), where "hedging" is "an evasive statement to avoid the risk of commitment" (Hall, 2000). Commonly used hedge phrases include *cannot exclude* and *not ruled out*.

From the perspective of annotation, hedging can be seen as ambiguity introduced by radiologists intentionally to keep certain information vague in order to protect themselves from potential lawsuits. If that information is related to what is being annotated, that could lead to annotation disagreement as annotators might interpret the radiologists' intention differently. As an example, one annotator labeled the sentence "*If clinically indicated, pelvic ultrasound could be performed in 4 to 6 weeks to document resolution*" as critical recommendation, but the other annotator did not because he thought the author was hedging.

4.3 Effects on the Annotation Process

The differences discussed in the previous section affect the annotation process in several ways.

4.3.1 Roles of NLP Researchers

In a typical annotation project, NLP researchers often play a central role; they are team leaders, guideline designers, technical support staff, and users. They consult linguistic experts to write annotation guidelines; they hire and train annotators; they monitor inter-annotator agreement and re-train annotators; they build NLP systems to pre-process data to speed up annotation.

However, they play a more limited role in our clinical annotation projects because they lack the medical expertise to (1) design the task and write guidelines (e.g., what do the three labels for CPIS mean), (2) select relevant patient records, (3) select and train annotators, and (4) foresee potential legal ramifications. Those tasks often fall on the shoulders of physicians, who play the roles of domain experts, annotators, guideline designers, and sometimes users.

4.3.2 Guidelines

In all of our annotation projects, annotations are done by physicians. Physicians are not familiar with common practice of annotation, such as creating detailed annotation guidelines in advance and revising guidelines if necessary. They are accustomed to making decisions (e.g., reading ICU reports and determining whether a patient has pneumonia) based on their professional training. They might not believe that writing detailed guidelines is necessary, and even if they want to, turning

[4] Health Information Portability and Accountability Act (HIPAA), Section 164.514. Available at: http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/adminsimpregtext.pdf

their medical knowledge into annotation guidelines can be very difficult. As a result, physicians often start annotation with no or very few annotation guidelines.

4.3.3 Finding and Training Annotators

Compared to annotators in a typical annotation task, physicians are much more expensive. They also have a very busy schedule and can spend very little time on annotation. Therefore, the common practice of training and re-training annotators, having several annotators work together to resolve disagreement, and having each annotator annotate 20 or more hours per week is all but impossible.

Training and re-training annotators is also difficult because the disagreement between physicians could be due to different interpretations related to their medical training. For instance, the sentence "*Correlation with ultrasound is advised*" is considered to be a critical recommendation by the radiologist but not by the clinician. (Re-)training them would mean that they have to change some long-held practice.

4.3.4 Process of Expert Annotation

Compared to traditional annotation schema, the team for clinical annotation is smaller, as the physicians (p) play the roles of guideline designers, annotators, and domain experts, whereas NLP researchers (n) provide technical support and build NLP systems using the corpus. The common process is as follows:

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

One important lesson we have learned from our projects is that NLP researchers should get more involved in the process, especially in Steps 4 and 6, as demonstrated in the next section.

5. Strategies

While we have a lot of experience with annotation in the general domain and the biomedical domain, we had not worked on expert annotation previously. In this section, we summarize a few lessons we have learned from these projects.

5.1 Importance of Annotation Guidelines

It is well-known that having detailed annotation guidelines is crucial for training annotators and ensuring high inter-annotator agreement (IAA). But for expert annotation, the annotators, who are medical experts in our case, already know how to determine whether a patient has a certain illness as it is part of their routine job; furthermore, NLP researchers would not know how to train them since the annotators have more knowledge about the task. We therefore ask the question whether detailed annotation guidelines are still necessary at all, and how often physicians' judgments would agree with each other without the guidelines?

In order to answer the question, we obtained double annotation on all the 800 reports in corpus C1, and 100 of the 1344 reports in corpus C2.[5] For each corpus, we asked each annotator to do two rounds of annotation:

1. In the first round, there were no annotation guidelines other than the definition of critical recommendation for corpus C1, and the meaning of labels for corpus C2 (e.g., "2a" means "no suspicion of PNA"). Each annotator annotated the data independently from each other.
2. In the second round, the annotators went over the instances (an instance is a sentence in C1 and a report in C2) that received different labels in the first round and did the following:
    i. For C1, each annotator wrote a note to explain the rationale for his labeling; then he read the rationale written by the other annotator and relabeled the sentences if he agreed with the other annotator's rationale.
    ii. For C2, the two annotators discussed all the reports that received different labels and came up with a detailed set of guidelines (see Table 4 for the guidelines for CPIS). They then waited for a few days (so that they would be unlikely to remember the decisions on the 100 discussed reports) and re-annotated the reports based on the guidelines.

For the second round, we prefer (ii) over (i) as (ii) requires annotators to come up with detailed guidelines, which would be valuable when annotating new data, but we could not do that for corpus C1 due to the busy schedules of its two annotators.

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.

Table 4: Annotation guidelines for determining CPIS labels in the chest X-ray corpus

With the two rounds of double annotations, we can calculate inter-annotator agreement (IAA) for each round. The results are shown in Tables 5-7. There are several observations. First, the IAA is pretty low for the first

[5] We did not do double annotation for Corpus C3 because we could not find another physician for the annotation task.

round, especially for the PNA labels in Table 7. Second, going through the second round with either (i) or (ii) improves the IAA significantly. Third, for the PNA labeling, the agreement is still low, 85%, even after the second round. All these indicate that solely relying on physicians' medical training is not sufficient in achieving a high IAA; creating detailed annotation guidelines and/or discussing examples with conflicting labels must be performed by physicians.

Round	A1	A2	Agreed	P/R/F	Kappa
1st	110	109	83	0.755/0.761/0.758	0.757
2nd	114	118	113	0.991/0.958/0.974	0.974

Table 5: 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 (i.e., positive sentences) marked by the annotators; 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.

Round	A1	A2	Agreed	Acc	kappa
1st	13/59/28	15/74/11	12/52/6	70%	0.415
2nd	13/72/15	16/72/12	13/68/10	91%	0.797

Table 6: IAA on **CPIS** 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 label from the two annotators; "kappa" is the kappa coefficient.

Round	A1	A2	Agreed	Acc	kappa
1st	44/32/24	69/26/5	36/5/4	45%	0.085
2nd	67/19/15	67/32/1	66/18/1	85%	0.697

Table 7: IAA on **PNA** labels for the 100 double annotated reports in the chest X-ray corpus (C2). x/y/z in a cell of the "A1", "A2", and "Agreed" columns are the numbers of reports with labels 2a, 2b, and 2c, respectively; "Acc" is the percentage of reports that receive the same PNA label; "kappa" is the kappa coefficient.

5.2 Providing Additional Information

Another lesson we learned from this experience is that, in addition to the label of the instance, we should also ask annotators to mark additional information such as evidence or rationale. For instance, Corpus C3 currently includes only 426 yes/no labels, one for each patient. We do not know what kind of evidence the annotator has found in the reports to support her decision, and which reports the evidence comes from. Ideally, we would prefer to have the annotator mark the evidence in the report (e.g., the text "WBC: 15000 mcl" in the discharge summary) and link it to the characteristics of PNA listed in Table 2 (e.g., "Abnormal white blood cell count" under "Clinical Signs and Symptoms"). Marking such information will not only help NLP researchers to build better systems (e.g., the systems can learn what kinds of cues are relevant to the class label), but also help annotators to resolve any annotation disagreement.

When choosing granularity of annotation, one always need to consider the benefits of fine-grained annotation vs. the downside of increased annotation time. For corpus C3, in order to give correct PNA labels, the annotators have to read the whole reports and look for those cues; as such, highlighting relevant text spans and clicking some buttons to link cues to some pre-defined characteristics would not substantially increase annotation time. The additional time is well spent since a patient has tens to hundreds of ICU reports, and therefore knowing where the cues come from will greatly reduce the number of features that an NLP classifier has to consider. We plan to include such additional information in the next stage of the project. For corpus C2, we also plan to mark the text span, although the benefits are less than in C3, because the reports in C2 are much shorter and the annotation is already at the report level, not the patient level.

5.3 Time Commitment from Physicians

All the projects discussed in Section 3 were initiated by our physicians. They are very interested in building NLP systems to meet their clinical needs. However, because they are not familiar with annotation process, they often underestimate the amount of time required for annotation, guideline designs, and other related activities. Their busy schedule at the hospital often limits the amount of time they can spend on the project.

To address this problem, we, the NLP researchers, should explain to the physicians what the annotation process looks like and why having detailed annotation guidelines and monitoring IAA are important. We should also provide them a good estimate of time commitment that will be required to complete the project. They can then make an informative decision on whether they are able to devote enough time to the project.

5.4 Early Involvement of NLP Researchers

Although it may be true that NLP researchers play a minor role in expert annotation, they should still get involved in the annotation process as early as possible. Despite their lack of medical training, they can help physicians in each step of the annotation process described in Section 4.3.4. For instance, they can calculate IAA and convince physicians to write detailed guidelines; they can inform physicians what kind of additional information would be beneficial to add; they can help physicians to decide how big the corpus needs to be; they can pre-process the data to filter out noisy data.

6. Conclusion

In this paper, we discuss three corpora that we created for clinical NLP projects. Unlike most of the previous annotation projects, these corpora require expert annotation. Our studies show that, without detailed guidelines and/or discussion, the annotation agreement among experts is low, indicating medical training itself is not sufficient for high-quality annotation. Although NLP researchers lack medical training and therefore play a minor role in guideline designs and annotation, their early involvement is important for the success of annotation.

7. References

Chapman W.W., Cohen K.B. (2009). Current issues in biomedical text mining and natural language processing. Journal of Biomedical Informatics. 42(5):757-759.

Chapman W.W., Nadkarni P.M., Hirschman L., D'Avolio L.W., Savova G.K., Uzuner O. (2011). Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. JAMIA[6]. 18:540-543.

Fiszman M., Chapman W.W., Aronsky D., Evans R.S., Haug P.J. (2000). Automatic detection of acute bacterial pneumonia from chest x-ray reports. JAMIA, 7(6):593-604.

Franzen K., Eriksson G., Olsson F., Lidin L.A.P., Coster J. (2002). Protein names and how to find them. International Journal of Medical Informatics. 67(1-3):49-61.

Hajic, J. (1999). Building a syntactically annotated corpus: The Prague Dependency Treebank. In E. Hajicova (Ed.), Issues of valency and meaning. Studies in honour of Jarmila Panevova. Prague, Czech Republic: Charles University Press.

Hall F. (2000). Language of the radiology report: primer for residents and wayward radiologists. AJR American Journal of Roentgenol. 175:1239-1242.

Kim J.D., Ohta T., Tateisi Y., Tsujii J. (2003). GENIA corpus–semantically annotated corpus for bio-text mining. Bioinformatics. 19:Suppl 1:180-2.

Kulick S., Bies A., Liberman M., Mandel M., McDonald R., Palmer M, Schein, Ungar L. (2004). Integrated annotation for biomedical information extraction. In Proceedings of HLT-NAACL workshop BioLink 2004, Linking Biomedical Literature, Ontologies, and Databases. pp. 61-8.

Levinson W. (1994). Physician-patient communication: a key to malpractice prevention. Journal of the American Medical Association (JAMA). 272:1619-1620.

Lucey L.L., Kushner D.C. (2010). The ACR Guideline on Communication: To Be or Not to Be, That Is the Question. Journal of the American College of Radiology. 7(2): 109-114.

Maamouri M and Bies A. (2004). Developing an Arabic treebank: methods, guidelines, procedures, and tools. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages.

Marcus M., Marcinkiewicz M.A., and Santorini B. (1993). Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, 19(2).

Miltsakaki E., Prasad R., Joshi A., and Webber B. (2004). The Penn Discourse TreeBank. In Proc. of LREC.

Pakhomov S.V., Coden A., Chute C.G. (2006). Developing a corpus of clinical notes manually annotated for part of speech. International Journal of Medical Informatics. 75(6):418-429.

Palmer M., Gildea D., and Kingsbury P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics. 31(1): 71-106.

Roberts A., Gaizauskas R., Hepple M., Davis N., Demetriou G., Guo Y., Kola J., Roberts I., Setzer A., Tapuria A., Wheeldin B. (2007). The CLEF Corpus: Semantic Annotation of Clinical Text. In Proceedings of AMIA Annual Symposium. pp. 625-629.

Roberts A., Gaizauskas R., Hepple M., Demetriou G., Guo Y., Roberts I., and Setzer A. (2009) Building a semantically annotated corpus of clinical texts. Journal of Biomedical Informatics. 42: 950-966.

Savova G.K., Chapman W.W., Zheng J., Crowley R.S. (2011). Anaphoric relations in the clinical narrative: corpus creation. JAMIA. 18:459-465.

Snow R., O'Connor B., Jurafsky D., Ng A.Y. (2008). Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In Proceedings of EMNLP'08. pp.254-263.

South B.R., Shen S., Jones M., Garvin J., Samore M.H., Chapman W.W., and Gundlapalli A.V. (2009). Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC Bioinformatics. 10(Suppl 9):S12.

South B.R., Shen S., Barrus R., DuVall S.L., Uzuner O., Weir C. (2011). Qualitative analysis of workflow modifications used to generate the reference standard for the 2010 i2b2/VA challenge. In Proceedings of the AMIA Annual Symposium. pp. 1243-1251.

Tanabe L., Xie N., Thom L.H., Matten W., Wilbur W.J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics. 2005;6(Suppl. 1):S3.

Towbin A.J., Hall S., Moskovitz J., Johnson N.D., Donnelly L.F. (2011) Creating a comprehensive customer service program to help convey critical and acute results of radiology studies. AJR American Journal of Roentgenology. 196(1):W48-51.

Uzuner O., Goldstein I., Luo Y., Kohane I. (2008) Identifying patient smoking status from medical discharge records. JAMIA. 15(1):14-24.

Uzuner O. (2009). Recognizing obesity and comorbidities in sparse data. Journal of the American Medical Informatics Association. 16(4):561-570.

Uzuner O, Solti I, Xia F., and Cadag E. (2010). Community Annotation Experiment for Ground Truth Generation for the i2b2 Medication Challenge. JAMIA. 17:519-523.

Uzuner O, South B.R., Shen S., DuVall S.L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. JAMIA. 18(5): 552-556.

Yetisgen-Yildiz M, Solti I., Xia F., Halgrim S.R. (2010) Preliminary Experiments with Amazon's Mechanical Turk for Annotating Medical Named Entities. In Proc. of Creating Speech and Language Data with Amazon's Mechanical Turk Workshop of NAACL'2010.

Yetisgen-Yildiz M, Gunn ML, Xia F, Payne T. (2011a). Automatic identification of critical follow-up recommendation sentences in radiology reports. In Proceedings of AMIA Annual Symposium. pp. 1593-1602.

Yetisgen-Yildiz M, Glavan BJ, Xia F, Vanderwende L, Wurfel MM. (2011b). Identifying Patients with Pneumonia from Free-Text Intensive Care Unit Reports. In Proceedings of Learning from Unstructured Clinical Text Workshop of ICML'2011.

Wallis A., McCoubrie P. (2011). The radiology report – Are we getting the message across? Clinical Radiology. 66(11):1015-1022.

Xia F., Palmer M., Xue N., Okurowski M.E., Kovarik J., Huang S., Kroch T., and Marcus M. (2000). Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. In Proceedings of LREC.

[6] *JAMIA* stands for Journal of American Medical Informatics Association.