Staggered NLP-assisted refinement for Clinical Annotations of Chronic Disease Events

Stephen T. Wu[1,2], Chung-Il Wi[1], Sunghwan Sohn[1], Hongfang Liu[1], Young J. Juhn[1]
[1]Mayo Clinic, Rochester, MN; [2]Oregon Health & Science Univ., Portland, OR
E-mail: wu.stephen@mayo.edu

Abstract

Domain-specific annotations for NLP are often centered on real-world applications of text, and incorrect annotations may be particularly unacceptable. In medical text, the process of manual chart review (of a patient's medical record) is error-prone due to its complexity. We propose a staggered NLP-assisted approach to the refinement of clinical annotations, an interactive process that allows initial human judgments to be verified or falsified by means of comparison with an improving NLP system. We show on our internal Asthma Timelines dataset that this approach improves the quality of the human-produced clinical annotations.

Keywords: annotation refinement, clinical text, staggered approaches, rule-based NLP

1. Introduction

Linguistic annotations for core Natural Language Processing (NLP) tasks – like POS tagging or parsing – are inherently analytic *explanations* of text. The purpose of a POS tag is to describe the grammatical structure of words in an utterance, and a corresponding resource aims to help computational methods reduce ambiguity in what is *communicated* by the text. In contrast, domain-specific annotations for NLP are often centered on real-world *applications* of text. A common reason to annotate events in medical or legal documentation is to highlight actionable information; this kind of language resource aims to help computational methods reduce disagreement in what is *decided* based on the text.

In this work, we describe the construction of Asthma Timelines, a clinically actionable resource in which chronic disease symptoms and events are annotated over clinical text. More specifically, with the help of medical experts, we annotated the prototypical chronic disease of asthma with its natural history (timeline of events) and relevant symptoms. These annotations span over multiple documents per patient, and are intended to provide evidence for a patient-level medical decision: is the patient's asthma status negative, positive, (in subsequent) remission, or (has it progressed to) relapse?

We observe that the challenging clinical nature of this problem makes gold standard annotations difficult to construct. Patients' asthma symptoms are primarily found in clinical text, which requires either NLP or human chart review to utilize; lab tests and other structured data fields are insufficient. While previous annotation projects (doing medical chart review) have reported reasonable agreement scores following a chart review guideline, the small amounts of associated data would be considered noisy training data for a machine learning algorithm. There are multiple clinical guidelines for what constitutes asthma (Yunginger et al. 1992a), and an annotator may not always follow one guideline rigorously. Also, while a human annotator can take context into account, there are times when this incorporation of contextual information can lead to *inconsistent* or even *incorrect* annotations. The typical means of overcoming noise in human annotations – using multiple annotators and measuring inter-annotator agreement – is a significant barrier when expert clinical annotators are necessary. These challenges are especially acute in privacy-protected domain-specific language technologies, and are representative of the inherent issues in evaluations rely solely on task-specific human annotations as a gold standard (Belz 2009).

Inspired by these challenges, we employ *staggered NLP-assisted refinement* to improve the quality and consistency of gold standard clinical annotations. We start with clinical guidelines, then stagger improvements between (a) having a medical expert provide/refine clinical annotations, and (b) having an NLP system classify patients (D'Avolio et al. 2010). While requiring a significant amount of effort, this approach provides medical experts an algorithmic means of assessing their adherence to a guideline, which, as we will show, improves the quality of the resulting annotations. This aligns with a shift away from a one-shot (frequently unfalsifiable) approach to annotation, and towards the efficiency of aided/adaptive annotation approaches, especially in specialized domains. We show that the process concretely improves the accuracy of manual annotations in Asthma Timelines.

1.1 Setting: Timelines of asthma status

Our work focuses on chronic diseases due to their prevalence and temporally evolving nature. It is estimated that almost one-half of Americans suffer from chronic diseases, but a patient's true disease status will progress over time in a way that may not be reflected in the clinical diagnosis of that disease. Therefore, we aim to build a resource for determining chronic disease status that includes timing markers alongside the statuses themselves.

We selected pediatric asthma for resource development because it is a prototypical chronic disease case, with a timeline that includes diagnosis, remission, and relapse. These asthma-related timeline events are not often reliably diagnosed in the course of routine medical care.[9] We believe that establishing a corpus of asthma timelines, annotated over language resources in EMR text, will allow future NLP systems to ascertain and track chronic disease status with accuracy and efficiency. Moreover, our annotations are actionable in the sense that they constitute data for public health research.

1.2 Setting: Population and EMR context

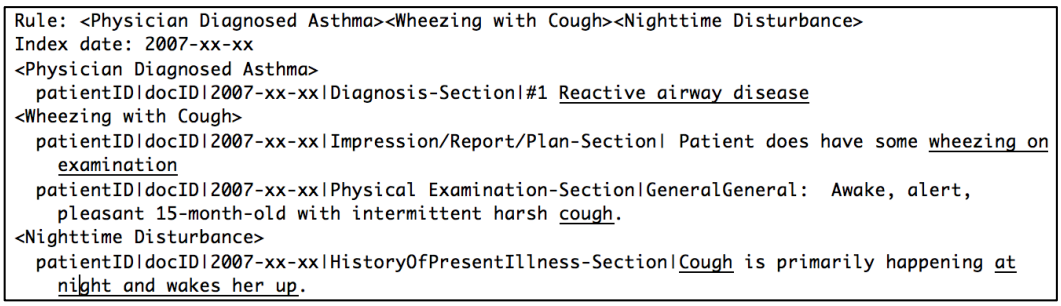

Figure 1: Visualization of the evidence for asthma status in the NLP system

Medical language (and decision-making derived from it) is highly dependent on the context in which the data is gathered. This context includes the characteristics of the underlying patient population that was sampled from the EMR, as well as what data was available from the implementation of the EMR. Medical records for these patients were primarily free-text documents written by health care providers (e.g., primary care doctors, specialists, nurses, and administrators). These documents include non-traditional text, like laboratory results. These were thoroughly reviewed and the outcomes of our interest were coded as binary values (e.g., asthma symptom yes/no) or dates (note date and event date).

We worked with EMRs for two groups of patients. The first group was a convenience-sample of the Mayo Clinic Sick Child Care cohort (Yoo et al. 2007) (n=115). Among these children, there were 35 who had positive asthma status (see below); the median age at asthma onset was 1.6 years. Of these 35, there were 17 children that progressed into remission; of the 17, 11 continued on to relapse. The second group consisted of 85 subjects who had positive asthma status from a larger cohort (2002-2006 Late Preterm Birth Cohort, n=542). The median age at asthma onset was 1.3 years. Among the Late Preterm Birth Cohort, there were 38 children that progressed into remission, including 11 with having subsequent relapse.

2. Methods

2.1 NLP System for asthma status

Algorithmic ascertainment of asthma status was accomplished by a rule-based NLP system implementing the Predetermined Asthma Criteria (PAC) (Yunginger et al. 1992b). We began with an algorithm reported in previous work (Wu et al. 2014; Wu et al. 2013), which we will call NLP_{v0}. In the course of the staggered annotation-refinement approach, we made a few updates to the system, resulting in NLP_{v1}-NLP_{v3}. The two most notable of these changes: we adopted the MedTagger (Liu et al. 2013) framework, and we provided detailed observational evidence – i.e., rules and context that contributed to the system's decision (see Figure 1). The visualization of this evidence allows human annotators to efficiently validate the results and promotes NLP-assisted annotation refinement.

2.2 Staggered NLP-assisted refinement of annotations

In staggered NLP-assisted refinement, we begin with an *initial annotation (I)* and an NLP algorithm; then, we stagger between arbitrarily ordered *system refinement steps (S)* and *annotation refinement steps (A)* until a termination condition is met.

2.2.1 Initial annotation of asthma status, symptoms, temporal expressions (I-step)

Our cohorts of patients had previously been annotated for asthma status (and its timing) according to the Predetermined Asthma Criteria; we will call this set of initial annotations PAC_{v1}. This identified each patient as being either positive or negative for asthma, at the time of any document in that patient's record.

In parallel, the rule-based NLP algorithm classified patients according to their asthma status, as described above.

We provided further temporal distinction to the evolution of a "positive" asthma status over time. Namely, we ascertained asthma remission/relapse for all patients who had some patient onset. We defined remission of asthma as lack of symptoms/signs of asthma or asthma-related medications or health care services for at least three consecutive years. Long-term remission was defined by no relapse of asthma after achieving remission. We also annotated ancillary asthma symptoms.

In this initial phase, we considered inter-annotator agreement on the manual chart review process by looking at a group of 15 patients. 100% of asthma statuses were in agreement, with both annotators estimating the timing of the statuses within 2 weeks of each other in each of the 15 cases.

2.2.2 System refinement (S-step)

When comparing I-step annotations to NLP output, the adjudicator discovered discrepancies that were problems in the rule-based NLP asthma ascertainment. This analysis of discrepancies provides knowledge that could be integrated into the rule-based NLP algorithm, so as to improve the consistency of the annotation process. This is similar to error analysis-based improvements for any rule-based system, and can be carried out multiple times in succession, or staggered after an annotation refinement step (A-step, below).

In our tests, using the 2002-2006 Late Preterm Birth Cohort, we examined all discrepant cases (false negatives and false positives) and updated the NLP system correspondingly. For example, we discovered that checking for a physician diagnosis of asthma in the "Final Diagnosis" section of a note helped reduce number of false positives for this variable. We

benchmarked improvements on this S-step twice (rows 2-3 in Table 1) before moving on. Our "staggered" approach also permitted later S-steps (row 5 in Table 1). Across the various S-steps, system errors mostly stemmed from inappropriate negated rules (e.g., denied wheezing, sister has asthma), cases where a history of asthma was self-reported but not confirmed by health care providers, or hypothetical situations (i.e., "if patient has wheezing").

2.2.3 Annotation refinement (A-step)

Because human judgments are often inconsistent, we allowed for comparisons of human annotations with system output to "falsify" the interim gold standard; thus, we had an adjudicator modify the human annotations in cases where the error was with human annotation rather than the NLP system.

On our data, the annotations for asthma status of 2002-2006 Late Preterm Cohort (n=542) were then compared with the NLP-based asthma status, resulting in 4-9% discrepancies (i.e., false positives and false negatives).

Analysis by an independent reviewer showed that the original annotator misclassified asthma status primarily due to overlooking key terms expressing symptoms that were part of the asthma criteria (e.g., wheezing episodes, night disturbance due to wheezing or cough). These discrepancies were adjudicated and the PAC_{v1} annotations on asthma status were modified accordingly. The results of this A-step are visible in row 4 of Table 1.

2.2.4 Termination

The staggered refinement approach continues until S-steps or A-steps do not improve the inter-annotator agreement, measured by Cohen's Kappa κ. This prevents adjudicators from overfitting the annotations to a particular system, and from unnecessary equivocation on inherently ambiguous instances.

3. Results and Discussion

Table 1 shows the metrics associated with both pairs of annotations (manual/expert and system/NLP) at each stage of staggered NLP-assisted refinement. Since we are *creating* the gold standard, the typical metrics of recall and precision for the NLP system are only measuring the adherence to the annotations up to that point. We can reverse this, and view the algorithm as the faithful implementation of annotation guidelines; "human recall" and "human precision" metrics are the simply the same values as "system precision" and "system recall," respectively.

The application of the 1st S-step appears to decrease system precision in favor of allowing some recall (row 2). This is due to the fact that the S-step is where new rules in the system will posit more (and perhaps inaccurate) matches. However, iterative use of the S-step (S_2-step, row 3) is able to overcome these limitations. With a more data-informed system (NLP_{v2}, row 3-4), the A-step reports sizable gains in all metrics. This shows that the initial NLP algorithm's rules did not have complete coverage of the needs and context of the user. The A-step makes a solid improvement on all metrics,

	Hum	Sys	Sys Rec/ Hum Prec	Sys Prec/ Hum Rec	F_1-score	Kappa κ
I-step	PAC_{v1}	NLP_{v0}	0.872	0.682	0.765	0.714
S_1 step	PAC_{v1}	NLP_{v1}	0.860	0.725	0.787	0.743
S_2 step	PAC_{v1}	NLP_{v2}	0.919	0.745	0.823	0.785
A step	PAC_{v2}	NLP_{v2}	0.940	0.887	0.913	0.892
S_3 step	PAC_{v2}	NLP_{v3}	0.960	0.865	0.910	0.888

Table 1: Staggered NLP-assisted refinement of annotations for Predetermined Asthma Criteria (PAC), comparing human (Hum) and NLP system (Sys) output

showing that the iterative process of an S-step has uncovered additional noise in the manual annotations. The final S_3-step is shown to not improve results on any metric, suggesting that the finite number of errors in human annotation may have been largely corrected.

The end result is a set of annotations that have very solid agreement with a rigorous, rule-based application of medical criteria. Moreover, these annotations are less noisy than those produced by two human annotators without the benefit of the staggered NLP-aided process.

4. Conclusion

We have described the creation of a corpus of clinical text, Asthma Timelines, annotated for asthma status and other supporting information. Crucial to the final form of the annotations was the process of staggered NLP-assisted refinement. We found that a rule-based classification system still needed tuning, and that a single pass on annotations would have left a significant amount of noise remaining in an imperfect "gold standard."

5. Acknowledgements

Thanks to Yanshan Wang for subsequent work with this data set. This work was supported in part by NIH grants R21AI116839 and R01AI112590.

6. Bibliographic References

Belz, Anja. 2009. 'That's nice... what can you do with it?', *Comput. Linguist.*, 35: 111-18.

D'Avolio, Leonard W., Thien M. Nguyen, Wildon R. Farwell, Yongming Chen, Felicia Fitzmeyer, Owen M Harris, and Louis D. Fiore. 2010. 'Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC)', *J Am Med Inform Assoc*, 17: 375-82.

Liu, H, SJ Bielinski, S Sohn, S Murphy, KB Wagholikar, SR Jonnalagadda, Ravikumar KE, ST Wu, IJ Kullo, and CG Chute. 2013. "An information extraction framework for cohort identification using electronic health records." In *AMIA Summits Transl Sci Proc*, 149-53. San Francisco, CA.

Wu, S. T., S. Sohn, K. E. Ravikumar, K. Wagholikar, S.

R. Jonnalagadda, H. Liu, and Y. J. Juhn. 2013. 'Automated chart review for asthma cohort identification using natural language processing: an exploratory study', *Annals of allergy, asthma & immunology : official publication of the American College of Allergy, Asthma, & Immunology*, 111: 364-9.

Wu, Stephen T, Young J Juhn, Sunghwan Sohn, and Hongfang Liu. 2014. 'Patient-level temporal aggregation for text-based asthma status ascertainment', *Journal of the American Medical Informatics Association*.

Yoo, K. H., S. K. Johnson, R. G. Voigt, L. J. Campeau, B. P. Yawn, and Y. J. Juhn. 2007. 'Characterization of asthma status by parent report and medical record review', *The Journal of allergy and clinical immunology*, 120: 1468-9.

Yunginger, J. W., C. E. Reed, E. J. O'Connell, L. J. Melton, 3rd, W. M. O'Fallon, and M. D. Silverstein. 1992a. 'A community-based study of the epidemiology of asthma. Incidence rates, 1964-1983', *Am Rev Respir Dis*, 146: 888-94.

———. 1992b. 'A community-based study of the epidemiology of asthma. Incidence rates, 1964-1983', *The American review of respiratory disease*, 146: 888-94.