Annotating Clinical Events in Text Snippets for Phenotype Detection

Prescott Klassen[1], Fei Xia[1], Lucy Vanderwende[2], and Meliha Yetisgen[1]
University of Washington[1], Microsoft Research[2]
PO Box 352425, Seattle, WA 98195, USA, One Microsoft Way, Redmond, WA 98052, USA
{klassp, fxia, melihay}@uw.edu, lucyv@microsoft.com

Abstract

Early detection and treatment of diseases that onset after a patient is admitted to a hospital, such as pneumonia, is critical to improving and reducing costs in healthcare. NLP systems that analyze the narrative data embedded in clinical artifacts such as x-ray reports can help support early detection. In this paper, we consider the importance of identifying the change of state for events—in particular, clinical events that measure and compare the multiple states of a patients health across time. We propose a schema for event annotation comprised of five fields ⟨location, attribute, value, change-of-state, reference⟩ and create preliminary annotation guidelines for annotators to apply the schema. We then train annotators, measure their performance, and finalize our guidelines. With the complete guidelines, we then annotate a corpus of snippets extracted from chest x-ray reports in order to integrate the annotations as a new source of features for classification tasks.

Keywords: Annotation, Event Extraction, Phenotype Detection

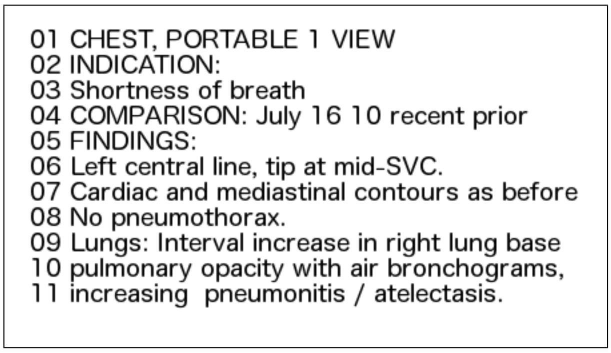
1. Introduction

A chest x-ray is a common type of medical report for monitoring the change in health of patients over time. Figure 1 presents an example chest x-ray. Our recent efforts have been to build an NLP system that analyzes the narrative embedded in x-ray reports to detect phenotypes/diseases such as pneumonia. To create the gold standard for training the system, we asked medical experts to annotate x-ray reports with phenotype labels and identify snippets of text that supported their labeling decision (e.g., Figure 1: lines 9-11). Analysis of the text snippets reveal that the snippets typically mention a change of state (COS) where a symptom is either initiating, increasing, persisting, decreasing, or terminating.

Monitoring the state of the patient, and comparing current state with previous states, is of critical importance to phenotype detection in the clinical scenario. In this paper, we propose to expand the annotation of COS to include the comparison of states over time. We implement a tuple schema of five fields ⟨location, attribute, value, change-of-state, reference⟩, create annotation guidelines and measure their performance by comparing the pairwise inter-annotator agreement between three annotators, and finalize our guidelines. Once completed, the guidelines and schema are applied to a corpus of text snippets extracted from chest x-ray reports. The resulting annotations are intended to be integrated as a new source of features for classification tasks in a pneumonia detection system.

2. Previous work on Phenotype detection

Early detection and treatment of ventilator associated pneumonia (VAP) is important as it is the most common healthcare-associated infection in critically ill patients. Even short-term delays in appropriate treatment for patients with VAP are associated with high mortality rates, longer-term mechanical ventilation, and excessive hospital costs. Our research goal is to build NLP systems which assist healthcare practitioners in identifying patients who are developing critical illnesses (e.g., VAP).



```
01 CHEST, PORTABLE 1 VIEW
02 INDICATION:
03 Shortness of breath
04 COMPARISON: July 16 10 recent prior
05 FINDINGS:
06 Left central line, tip at mid-SVC.
07 Cardiac and mediastinal contours as before
08 No pneumothorax.
09 Lungs: Interval increase in right lung base
10 pulmonary opacity with air bronchograms,
11 increasing  pneumonitis / atelectasis.
```

Figure 1: Sample chest x-ray report

2.1. PNA/CPIS Detection

In our previous study (Tepper et al., 2013), we built an NLP system to detect pneumonia, and the system was trained on an annotated corpus of 1344 chest x-rays from the UW Harborview Medical Center. Annotators read each report and determined whether the patient had pneumonia (PNA) and also recorded a clinical pulmonary infection score (CPIS) (Zilberberg and Shorr, 2010). CPIS is used to predict which patients will benefit from the invasive, and preferably avoidable procedure to obtain pulmonary cultures. There are three labels for CPIS *1A (no infiltrate)*, *1B (diffuse infiltrate or atelectasis)*, and *1C (local infiltrate)*. Similarly, there are three labels for PNA *2A (no suspicion of PNA)*, 2B (suspicion of PNA), and *2C (probable PNA)*. In addition to labels for CPIS and PNA, we also asked the annotators to highlight the text snippet in the chest x-rays that supports the labels that annotators choose for the x-rays. We call these snippets *rationale snippets* (see (Yu et al., 2011) for a similar approach).

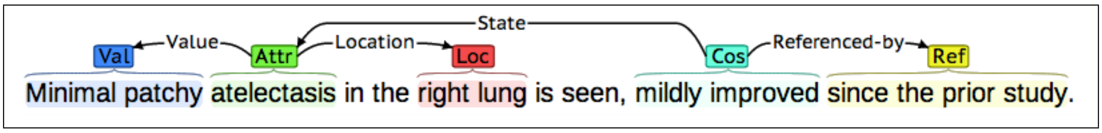

State
Val
Value
Attr
Location
Loc
Cos
Referenced-by
Ref
Minimal patchy atelectasis in the right lung is seen, mildly improved since the prior study.

Figure 2: A snippet featuring an event annotation connecting all five fields of the COS tuple.

2.2. Experiment Results

First, two sequence labelers were trained to predict the locations of CPIS and PNA rationale snippets, and the f-scores for predicting snippets are 78.7% (for PNA) and 93.9% (for CPIS), respectively. Next, two SVM classifiers, one for CPIS and the other for PNA, were trained and evaluated using 5-fold cross validation. Three sets of experiments were run using features selected from the whole document, the predicted snippets alone, or a combination of both. The same features were used for all three sets of experiments, including unigrams and bigrams filtered for stopwords, digits, and punctuation, UML concepts, and alternate proposing conjunctions (versus, or, etc.). Snippets alone performed the best for both classification tasks, resulting in a macro F1-score of 76.9% and an accuracy score of 87.1% for CPIS, and a macro F1-score of 74.0% and an accuracy score of 82.1% for PNA (Tepper et al., 2013).

Error analysis for both classification tasks reveals that resolving classification errors requires features that go beyond simple concepts or word ngrams. Consider the snippet *"The previously noted right upper lobe opacity consistent with right upper lobe collapse has resolved"*, which is labeled in the gold standard 1A (no infiltrate). The system mislabeled it 1C (localized infiltrate), because the snippet supports 1C entirely up until the crucial words "has resolved". These types of errors motivate the COS annotation task we describe in this paper.

3. Change of State for Clinical Events

Our corpus contains annotated radiology reports and highlighted snippets of text where annotators found support for their annotations. These snippets frequently describe observations of change and such COS observations appear more often in snippet text than in non-snippet text[1].

Previous COS analyses (e.g., (Sun et al., 2012)) have focused on an analysis where events are expressed as verbs. In our data, however, many of the events are expressed as nouns and adjectives. Our annotation scheme is explained below.

3.1. Description of COS tuple

The event analysis in (Uzuner et al., 2010; Uzuner et al., 2011; Albright et al., 2013) mark multiple types of events, temporal expressions, and event relations whereas our annotation is constrained to tracking changes in a patients medical condition. We define an event in our corpus as a tuple, consisting of the fields *loc*, *attr*, *val*, *cos*, and *ref*, where loc is the anatomical location (e.g., *right lung* in Figure 2), attr is an attribute of the location that the event is about (e.g., *atelectasis* in Figure 2), val is a possible value for the attribute (e.g., *minimal patchy* in Figure 2), cos indicates the change of state for the attribute value compared to some previous reports (e.g., *mildly improved* in Figure 2), and ref is a link to the report(s) that the change of state is compared to (e.g., *since the prior study* in Figure 2). Not all tuples will have values for all five fields. A field can be unspecified and inferred from the context of the surrounding snippet text or from the collection of snippets that have been extracted from the sequence of a patients x-ray reports. More discussion about the event definition can be found in (Vanderwende et al., 2013).

4. Change of State (COS) Annotation

In this section we describe our corpus of extracted text snippets, the tools and processes we implemented for annotation, and report on inter-annotator agreement.

4.1. Development Process

We annotated the events in the x-ray reports in several steps:

1. Extract text snippets from x-ray reports.
2. Create annotation guidelines.
3. Create an annotation tool.
4. Train annotators, conduct triple annotation on a small sample of the snippets, compare annotations, and finalize annotation guidelines.
5. Annotate the complete corpus, following the finalized annotation guidelines.

Steps 1, 3, and 4 are explained below.

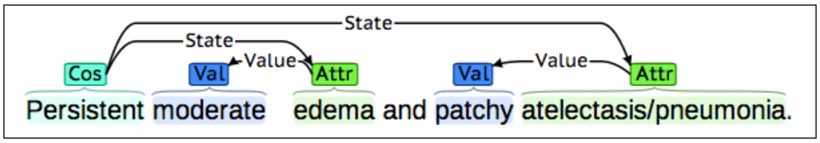
4.2. Annotation Tools and Schema

The event tuple structure described in Section 3.1 was translated into a schema for use with the BRAT[2] annotation tool (Stenetorp et al., 2012). BRAT is a lightweight, web browser-based annotation tool which is easy to install and use across major operating systems. BRAT allows annotators to mark entities and link entities by arcs.

To use BRAT for our annotation, we mark five types of entities, corresponding to the five fields in an event tuple. To mark an event tuple, because some fields are shared by multiple events, we decided to use directed, labeled arcs to link two entities. There are four arc labels: (1)The label *State* connects a Cos entity with an Attr, Loc, or Val entity; (2) *Value* connects an Attr or Loc entity with a Val entity;

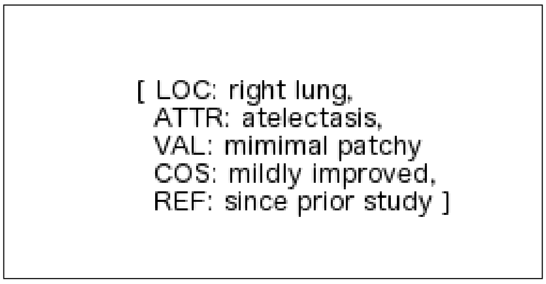
[1] Out of a random sample of 100 snippets and 100 non-snippet texts, 83 snippets contain mentions of COS; in contrast, 61 non-snippet texts contain mentions of COS.

[2] http://brat.nlplab.org/



Persistent moderate edema and patchy atelectasis/pneumonia.

Figure 3: A snippet featuring shared entities between events.



[LOC: right lung,
ATTR: atelectasis,
VAL: mimimal patchy
COS: mildly improved,
REF: since prior study]

Figure 4: An event tuple extracted from the graph in Figure 2.

(3) *Location* connects an Attr or Loc entity with a Loc entity, and (4) *Referenced-by* connects an Attr or Cos entity with a Ref entity.

Once the entities in a snippet are connected to one another by labeled arcs, one or more connected, directed graphs are formed by these arcs. See Figure 2 for a screen capture of our schema applied to a text snippet. From the graph, event tuples are generated. Figure 4 shows the event tuple generated from the graph in Figure 2.

Note that events can share entities between them. The graph in Figure 3 features two attr entities connected to a single change of state entity. Two event tuples are generated from the graph, both featuring the same shared change of state entity.

In addition to the description of the annotation tool and schema, the annotation guidelines include a list of common terms from the corpus and their mappings to entities in the schema.

4.3. Snippet Corpus Description

As mentioned in Section 2.1, our corpus consists of 1008 unique text snippets extracted from 1344 x-ray reports. The snippets were selected by medical experts to support their decision when annotating the reports with PNA/CPIS labels. Table 1 provides some statistics of the corpus.

Corpus Item	Count
X-ray Reports	1344
Unique Snippets	1008
Entities	7173
Labeled Arcs	4128
Event Tuples	2101

Table 1: Statistics of the snippet corpus

4.4. Inter-annotator agreement

To train annotators, calculate the inter-annotator agreement, and finalize the annotation guidelines, 100 snippets were selected at random from our corpus of 1008 unique text snippets. They were annotated in two stages by three annotators. In the first stage, the annotators read the annotation guidelines, learned to use the annotation tools, and then annotated the first 20 snippets. The annotators then met and compared annotations, and the feedback from this discussion was used to revise and finalize annotation guidelines. In the second stage, the annotators revised their annotation of the first twenty snippets following the revised guidelines, and completed annotating the remaining 80 snippets.

To calculate inter-annotator agreement, we compare the annotations of each of the annotator pairs at three levels: word level, entity level, and event level. Each level provides a different aspect about annotation disagreements between annotators and contributes to the development of annotation guidelines.

4.4.1. At the word level

To compare annotation at the word level, we use the standard BIO scheme to obtain word-level labels from entity annotation. That is, if a text span is labeled as an entity of type X, the word-level label of the first word in the span is B-X, and the label of other words in the span is I-X, and words that do not appear in any entity has label O.

	Precision	Recall	F1-score
A/B	0.8607	0.8434	0.8520
A/C	0.8532	0.8579	0.8555
B/C	0.7862	0.8330	0.8089
Average	0.8334	0.8448	0.8388

Table 2: Inter-annotator agreement at the word level for the first 20 snippets. A, B, and C denote the three annotators. The scores are micro-averages over different word-level labels.

	Precision	Recall	F1-score
A/B	0.8181	0.8370	0.8274
A/C	0.8532	0.8579	0.8555
B/C	0.7983	0.7912	0.7947
Average	0.8232	0.8287	0.8259

Table 3: Inter-annotator agreement at the word level for the final 100 snippets.

With the word-level labels, we calculate precision, recall, and F1-scores of B-X and I-X labels in a pairwise comparison of three annotators for both the first 20 and final 100 snippets, and the results are in Tables 2-3.

4.4.2. At the entity level

To compare annotations at the entity level, we define exact match of entities. Two entities match exactly if their text spans are exactly the same and their entity types are identical.

Tables 4 and 5 show precision, recall, and F1-score in a pairwise comparison of three annotators for both the first 20 and final 100 snippets.

	Precision	Recall	F1-score
A/B	0.9037	0.8920	0.8965
A/C	0.7925	0.8127	0.8007
B/C	0.8237	0.8560	0.8367
Average	0.8399	0.8536	0.8446

Table 4: Inter-annotator agreement at the entity level for the first 20 snippets.

	Precision	Recall	F1-score
A/B	0.8732	0.8852	0.8787
A/C	0.8187	0.8318	0.8247
B/C	0.8447	0.8452	0.8445
Average	0.8455	0.8541	0.8493

Table 5: Inter-annotator agreement at the entity level for the final 100 snippets.

4.4.3. At the event level

From the labeled graph (See Figures 2 and 3), we designed an algorithm to derive event tuples. Two event tuples are considered an exact match if they have the same fields filled out and the entities for those fields match exactly. This is the strictest measure among the three, because two events would not match if one field in one event does not exactly match the same field in the other event.

Tables 6 and 7 show precision, recall, and F1-score in a pairwise comparison of three annotators for both the first 20 and final 100 snippets.

	Precision	Recall	F1-score
A/B	0.6829	0.6829	0.6829
A/C	0.5610	0.5349	0.5476
B/C	0.6098	0.5814	0.5952
Average	0.6179	0.5997	0.6086

Table 6: Inter-annotator agreement at the event tuple level for the first 20 snippets.

The inter-annotator agreements (F1-scores) at the word level and the entity level are about 0.82-0.85 for both the first 20 and the final 100 snippets, indicating that labeling entities is relatively easy. Lower scores for the final 100 in Table 3 is due to the occurrences of some new ambiguous words that were not seen in the initial 20 snippets. Examples include words like *parenchymal* labeled as part of a value label rather than a location label, or *airspace disease* labeled as an attribute or broken into a location (airspace) and attribute (disease). Differences in text span offsets boundaries led to general rules for labeling value entities as individual text spans and location entities as multi-word spans. For example, the words *minimal patchy* were separated into two value text spans whereas upper right lobe was combined into a single location text span. Similarly, the final 100 snippets contain some unseen ambiguous text spans, which contributed to a minimal improvement in inter-annotator agreements F1-scores. Examples of unseen ambiguous text spans in the final 100 included coordination construction such as *consolidation versus atelectasis* and *atelectasis, effusions, or consolidation*, which annotators either treated the whole string as a single text span or labeled each conjunct as an individual text span and did not labeled the coordinating conjunctions *versus* and *or* as part of an entity.

	Precision	Recall	F1-score
A/B	0.6798	0.6935	0.6866
A/C	0.7241	0.7067	0.7153
B/C	0.7638	0.7308	0.7469
Average	0.7226	0.7103	0.7163

Table 7: Inter-annotator agreement at the event tuple level for the final 100 snippets.

Agreement at the event level was lower than the other levels due to its exact match requirement: two event tuples match only if all entity fields match. However, the event tuple agreement for the final 100 snippets is higher than the first 20 because the discussion in the first stage helped to clarify for annotators how to annotate events. One specifically addressed issue was where to attach ambiguous entities to the graph, like reference, which could either attach to the overall change-of-state entity or its child attribute entities.

An analysis of the differences between annotators resulted in updated guidelines which were then followed by a single annotator who completed the annotation of the entire corpus of 1008 snippets. The annotated corpus is available at *http://depts.washington.edu/bionlp/datasets.htm.*

5. Conclusion and future work

General-domain event annotation without a target application can be challenging. Our annotation focuses on the marking of COS in medical reports because COS is an important indicator of the patients medical condition. We proposed a schema where an event is a $\langle$loc, attr, val, cos, ref$\rangle$ tuple, and annotated snippets extracted from x-ray reports. Our experiments showed strong agreement between three annotators at word, entity, and event levels.

For future work, we plan to use the corpus to train an event detector and then add event-based features to our phenotype detection system. We expect that such features will improve phenotype detection accuracy just as (Bejan et al., 2013) demonstrated that adding features that encode negation and assertion information improved phenotype classification accuracy. We will also extend our schema to annotate relations between events (e.g., one event causes another event).

Our ultimate goal is to use event detection, phenotype detection, and other NLP systems to monitor patents medical conditions over time and prompt physicians with early warning, and thus improve patient healthcare quality while reducing the cost of healthcare.

6. Acknowledgements

This project was partially funded by the University of Washington Research Royalty Fund and the Institute of Translational Health Sciences grant UL1TR000423. We would also like to thank the anonymous reviewers and Wen-wai Yim for helpful feedback.

7. References

Albright, D., Lanfranchi, A., Fredriksen, A., Styler IV, W. F., Warner, C., Hwang, J. D., Choi, J. D., Dligach, D., Nielsen, R. D., Martin, J., Ward, W., Palmer, M., and Savova, G. K. (2013). Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of American Medical Informatics Association (JAMIA)*, 20:922–930.

Bejan, C. A., Vanderwende, L., Xia, F., and Yetisgen-Yildiz, M. (2013). Assertion modeling and its role in clinical phenotype identification. *Journal of American Medical Informatics Association (JAMIA)*, 46(1):68–74.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T.-k., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations Session at EACL 2012*.

Sun, W., Rumshisky, A., Uzuner, O., Szolovits, P., and Pustejovsky, J. (2012). The 2012 i2b2 temporal relations challenge annotation guidelines. Manuscript, Available at https://www.i2b2.org/NLP/TemporalRelations/Call.php.

Tepper, M., Evans, H. L., Xia, F., and Yetisgen-Yildiz, M. (2013). Modeling annotator rationales with application to pneumonia classification. In *Expanding the Boundaries of Health Informations Using AI Workshop of AAAI 2013*.

Uzuner, Ö., Solti, I., Xia, F., and Cadag, E. (2010). Community annotation experiment for ground truth generation for the i2b2 medication challenge. *Journal of American Medical Informatics Association (JAMIA)*, 17(5):519–23.

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of American Medical Informatics Association (JAMIA)*, 18(5):552–6.

Yu, S., Farooq, F., Krishnapuram, B., and Rao, B.-r. (2011). Leveraging rich annotations to improve learning of medical concepts from clinical free text. In *ICML workshop on Learning from Unstructured Clinical Text*, Bellevue, WA.

Zilberberg, M. D. and Shorr, A. F. (2010). Ventilator-associated pneumonia: the clinical pulmonary infection score as a surrogate for diagnostics and outcome. *Clinical Infectious Diseases*, 51(Suppl 1):S131–S135.