

# Annotating Change of State for Clinical Events

## 1 Introduction

In this document, we layout guidelines for marking change of state (COS) annotations in x-ray reports as described in (Tepper et al., 2013; Klassen et al., 2014).

### Corpus Description

The corpus that accompanies this guideline document has been annotated with change of state annotations. It is a collection of 1065 de-identified rationale snippets (of which 1008 are unique) that have been extracted from a corpus of 1344 chest x-ray reports from the University of Washington Harborview Medical Center. The rationale snippets are one or more sentences or sentence fragments that annotators selected from an overall x-ray report because they provide a rationale for the classification of the report in regards to two measures, evidence of pneumonia (PNA) and clinical pulmonary infection score (CPIS). These two measures were used in (Tepper et al., 2013; Klassen et al., 2014) to predict ventilator acquired pneumonia (VAP).

The corpus is divided into individual text files which contain one hundred rationale snippets. The annotation tool, BRAT (<http://brat.nlplab.org/index.html>), was used to annotate the corpus and a separate associated annotation file in the BRAT annotation format contains the change of state annotations and their location in the text file by character start and end offsets. Two configuration files for BRAT are also included in the annotation directory, `annotation.conf`, which contains the schema definition of change of state annotations, and `visual.conf`, which contains colors and layout rules for the individual annotation elements.

### Stage 1 Annotation Process

COS annotations were created to mark change of state and diagnosis information in the rationale snippets. We mark change of state and diagnoses events in chest x-ray reports. This is accomplished in several stages. The current document focuses on Stage 1.

## 2 Decisions for Stage 1

In Stage 1, we make the following decisions in order to expedite the annotation process:

1. Annotate snippets only:

As our previous studies show (Tepper et al., 2013), extracting features from only text within rationale snippets for CPIS/PNA classification outperforms using features extracted from the text of the whole x-ray report. Given that snippets are more informative and only a small percentage of sentences are snippets we decide to annotate snippets only in Stage 1 of the change of state annotation process.

2. Annotate without context:

Snippets are annotated without looking at the larger context (e.g., the whole X-ray report or other reports for the same patient). This allows snippets to be annotated in a single file, instead of loading thousands of files, many of which do not contain snippets.

3. Concepts are NOT pre-identified in the snippets:

While one could run tools such as MetaMap to identify medical concepts as a preprocessing step, we decide not to do that in Stage 1 for several reasons: first, the quality of MetaMap in identifying medical concepts is questionable. Second, the relationship between a medical concept and the fields in a COS tuple is unclear at this stage. For instance, a field could correspond to part of a concept or multiple concepts.

4. Do not fill in the implicit values:

The values of some fields can be inferred based on the context one's medical knowledge; for instance, in the text "the lungs are clear", an attribute like "density" might apply. The annotators are NOT adding this field manually at Stage 1 because adding such info requires domain knowledge.

5. Do not do anaphor resolution:

A field could contain anaphor, e.g., "the left one is clear", where "one" refers to lung. Anaphor resolution could take time, and sometimes the referent is outside the snippet. So we decide to do anaphor resolution in later stages. There is one exception to this decision: see Ex 8 in Section 3.3 and Part 4 of Appendix.

### 3 Annotation

In Stage 1, we annotate two types of events: change of state and diagnosis.

#### Change of state event

A change of state event is represented as a tuple (**loc**, **attr**, **val**, **cos**, **ref**). Every field is optional, but all COS tuples must contain either a **val**, **attr**, or **cos** field. The five fields are defined as follows:

- **loc**: anatomical location (e.g., lung).
- **attr**: something doctors are measuring or observing (e.g., volume, opacity) including symptoms or diseases (e.g., edema).
- **val**: a possible value for an attr (e.g., clear). When an attr is a symptom/disease, the val can be a string that indicates presence or absence (e.g., “there is”).
- **cos**: change of state compared to other reports for the same patient (e.g., unchanged)
- **ref**: a link to the mention of a previous report(s) or observation that the change of state is being compared to (e.g., prior examination)

Change of state tuples are automatically generated from change of state annotations, which are labeled text spans (entities) connected by directed labeled arcs (relations). The names of the labeled text spans (entities) in our change of state annotations correspond to the labels defined in our event tuple definition: **loc**, **attr**, **val**, **cos**, and **ref**. An additional entity, **conj**, is used to associate multiple entities in aggregate when connecting entities in an event tree or graph.

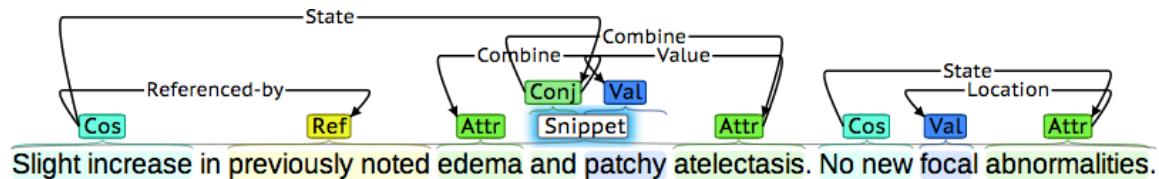
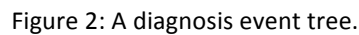


Figure 1: An example sentence marked with change of state annotations using the BRAT tool.

In Figure 1 above, the rationale snippet contains two sentences and two event trees. The change of state event tuples automatically extracted from the two event trees are:

[**loc**: -, **attr**: edema, **val**: -, **cos**: slight increase, **ref**: previously noted]  
[**loc**: -, **attr**: atelectasis, **val**: patchy, **cos**: slight increase, **ref**: previously noted]  
[**loc**: -, **attr**: abnormalities, **val**: focal, **cos**: No new, **ref**: -]

Very often a report will include a ‘diagnosis’ statement made by physicians; they will often, but not always, give “differential” diagnoses. Such information is very useful for phenotype detection. The statement may include a hedge or other assertion types. In Stage 1, we mark the whole text chunk as a diagnosis, as well as labeled text spans similar to change of state events, but headed by a diagnosis head (Dhead) instead of a cos entity. Diagnosis event tuples can be generated in similar way to the change of state, but instead of a cos field, a diagnosis tuple has a dhead field.



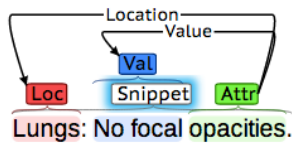
[loc: -, attr: edema, val: -, dhead: likely, ref: -]  
[loc: -, attr: infection, val: -, dhead: likely, ref: -]

Below are some examples with both change of state and diagnosis event annotations as labeled text spans connected by labeled directed arcs as an event tree and as a tuple with change of state or diagnosis event fields:

- 

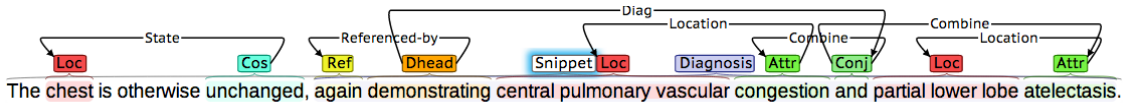
[**loc:** *lungs*, **attr:** -, **val:** *clear*, **cos:** -, **ref:** -]

2. *Lungs: No focal opacities.*



[loc: lungs, attr: opacities, val: No focal, cos: -, ref: -]

3. *The chest is otherwise unchanged, again demonstrating central pulmonary vascular congestion and partial lower lobe atelectasis.*

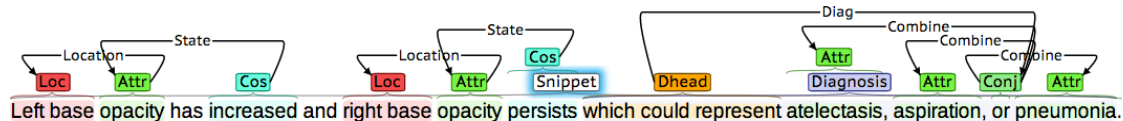


[loc: chest, attr: -, val: -, cos: unchanged, ref: -]

[loc: central pulmonary vascular, attr: congestion, val: -, dhead: demonstrating, ref: again]

[loc: partial lower lobe, attr: atelectasis, val: -, dhead: demonstrating, ref: again]

4. *Left base opacity has increased and right base opacity persists which could represent atelectasis, aspiration, or pneumonia.*



[loc: left base, attr: opacity, val: -, cos: increased, ref: -]

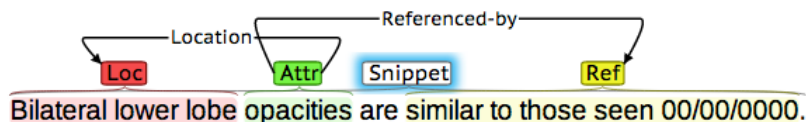
[loc: right base, attr: opacity, val: -, cos: persists, ref: -]

[loc: -, attr: atelectasis, val: -, dhead: which could represent, ref: -]

[loc: -, attr: aspiration, val: -, dhead: which could represent, ref: -]

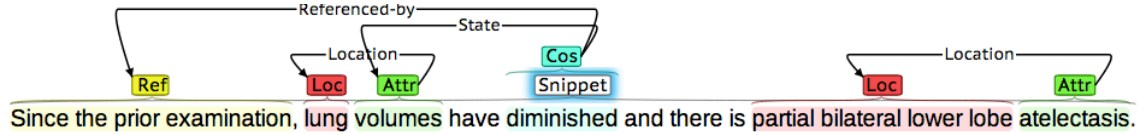
[loc: -, attr: pneumonia, val: -, dhead: which could represent, ref: -]

5. *Bilateral lower lobe opacities are similar to those seen on DATE.*



[loc: Bilateral lower lobe, attr: opacities, val: -, cos: -, ref: similar to those seen 00/00/0000]

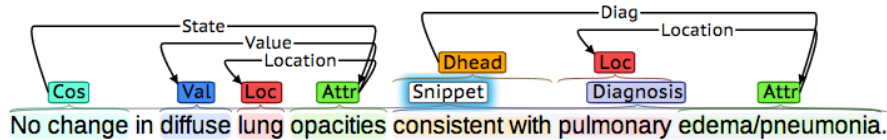
6. Since the prior examination lung volumes have diminished.



[loc: lung, attr: volumes, val: -, cos: diminished, ref: since the prior examination]

[loc: partial bilateral lower lobe, attr: atelectasis, val: -, cos: -, ref: -]

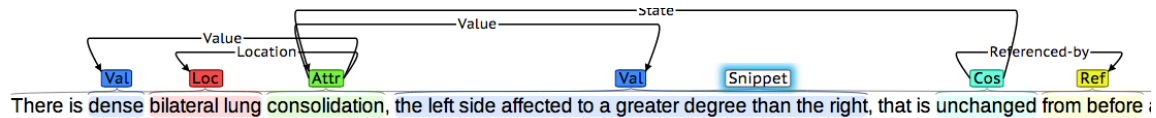
7. No change in diffuse lung opacities consistent with pulmonary edema/pneumonia.



[loc: lung, attr: opacities, val: diffuse, cos: no change, ref: -]

[loc: pulmonary, attr: edema/pneumonia, val: -, dhead: consistent with, ref: -]

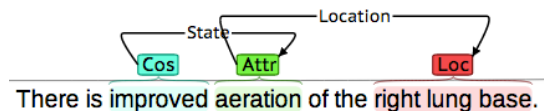
8. There is dense bilateral lung consolidation, the left side affected to a greater degree than the right, that is unchanged from before



[loc: bilateral lung, attr: consolidation, val: dense, cos: unchanged, ref: from before]

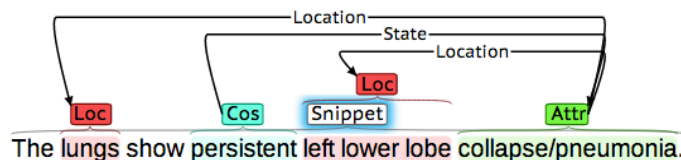
[loc: bilateral lung, attr: consolidation, val: the left side affected to a greater degree than the right, cos: unchanged, ref: from before]

9. There is improved aeration of the right lung base



[loc: right lung base, attr: aeration, val: -, cos: improved, ref: -]

10. The lungs show persistent left lower lobe collapse/pneumonia



[loc: lungs, attr: collapse/pneumonia, val: -, cos: persistent, ref: -]

[loc: left lower lobe, attr: collapse/pneumonia, val: -, cos: persistent, ref: -]

## 4 Preprocessing steps

The x-ray reports are preprocessed before loading into the annotation tool and presented to annotators:

- a. Snippets are extracted from the reports and combined into text files of 100 snippets. Snippets are delimited by line in the text file. The offsets of the snippets in the original reports are associated with a snippet entity in a BRAT annotation file. The offsets of the text span are stored along with its label in a tab and space delimited text format.
- b. An attempt to de-duplicate snippets if they repeat is made to maximize the efficiency of the annotation process. This de-duplication was not done in our example corpus.

We will not run sentence segmentation and tokenization because adding whitespace to the snippets will complicate the task of recovering the offsets in the original reports.

## 5 Using the BRAT annotation tool

To install the annotation tool, follow the instructions for your operating systems as described in the BRAT installation instructions:

<http://brat.nlplab.org/installation.html>

To understand how to annotate a document in BRAT, follow the simple tutorial at:

<http://brat.nlplab.org/introduction.html>

Our snippet corpus is divided into collections of 100 snippets per file. Load the example corpus into your brat data directory, navigate to the collection, and open the first set of annotations in file snippet1\_100 in your browser. Make sure to login to BRAT in order to edit or add annotations.

To annotate entities, simply select a text span with your mouse and options to label the text spans will appear in a modal dialog box. Select the appropriate label and click 'OK'.

To connect entities with labeled arcs, select an entity with your mouse and drag your mouse to another entity, a dialog box for the type of labeled arc or relation will appear. Select the appropriate label and click 'OK'.

To modify a label, simply double click an existing entity and a modal dialog will appear, allowing you to change or delete a label for a text span (entity) or a labeled arc (relation).



## 6 Appendix A: Common questions for annotation

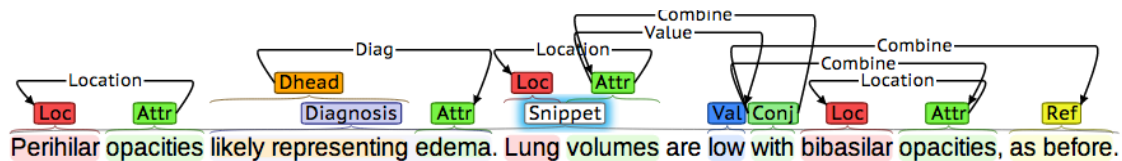
We divide common questions into several groups.

### Part 1: What kind of events should be annotated?

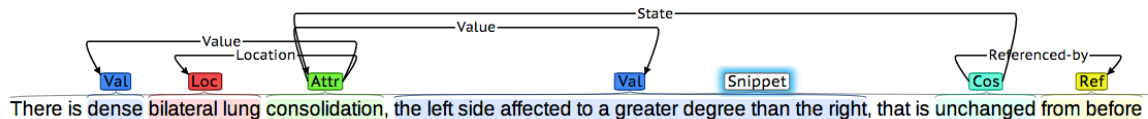
There are change of state and diagnosis event trees, consisting of labeled text spans and directed labeled arcs, which can be minimally represented with a sparse tuple. These events may have implied values for their fields which are not marked in Stage 1:

- Minimal event trees can be as sparse as an **attr** and a **loc** ('Perihilar opacities' in the figure below).
- Maximal event trees may contain many different instances of an event tuple.

We use the **conj** entity and **combine** relation to attach labeled text spans to aggregate change of state labels. For example, in the sentence below:



Values may be significant phrases rather than simple adjectives, for example, In "There is dense bilateral lung consolidation, the left side affected to a greater degree than the right, that is unchanged from before and in keeping with bilateral pneumonia": the underlined text is labeled as a **val** of the **attr** *consolidation*.



## Part 2: What is the span of a field?

Q1: How big should the span be? E.g., should modifiers be included?

A1: The span should include all and only the words (including modifiers) that are needed to (uniquely) identify the field at the appropriate scope. A span does not need to be a full XP (e.g., NP) or contiguous.

- Adverb:

In “the lungs are otherwise clear”, should “otherwise” be part of **val**?

No, the **val** is “clear”. Because “otherwise” implies a comparison with a value that is not present in the sentence.

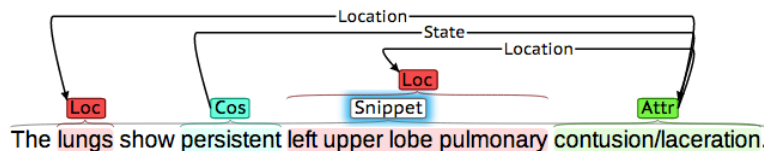
- Determiner:

In “the lung is clear”, should “the” be part of **loc**?

No. Do not include “the” because we will attempt to label entities as minimally as possible.

Q2: Can the span for a field be noncontiguous? If so, how to mark it?

A2: Yes. For instance, in “The lungs show persistent left lower lobe collapse/pneumonia”, the **loc** is “the lungs ... left lower lobe”. The annotation tool allows for multiple noncontiguous values of a field ; just mark each part of the span individually, and give them the same field name.



Q3: Can a span appear in multiple events?

A3: Yes, e.g., “unchanged” in “both edema and atelectasis are unchanged” appears in two event tuples as **cos**.

### Part 3: Which field is it?

In this section, we provide criteria for distinguishing confusing field pairs.

Q1: **attr** vs. **val**?

A1: **attr** is something that we are measuring or observing (which includes symptom/disease); **val** is measurement of the **attr**;

Q2: **val** vs. **cos**?

A2: If the field is an absolute value, with no comparison, it is **val**. When the field indicates comparison, if the comparison is with the previous report of the same patient, it is **cos**. If it is with other patients or people, it is **val**. Sometimes, distinguishing the two requires domain knowledge.

- “No focal abnormalities”: “focal” is **val**, “abnormalities” is **attr**, “No” is **val** as it indicates the absence of the symptom.
- “No new focal abnormalities”: “focal” is **val**, “abnormalities” is **attr**, “no new” is **cos**. The reason is that “new” indicates the comparison with previous reports.
- “The previously noted right upper lobe collapse has now reexpanded”: “right upper lobe” is **loc**, “collapse” is **attr**, “now reexpanded” is **cos** “Previously noted” is **ref**.

### Part 4: How should a construction be handled?

**Negation:**

Q1: Should negated words (e.g., “not” in “the lungs have not been clear”) be included in a span?

A1: There are two ways to deal with negation:

- (1) to include the negated word in a field
- (2) to add a separate flag for negation and then link it to the field that is negated

Option (1) is easier to annotate and it is easy to infer (2) from (1). So we choose (1). This option could lead to noncontiguous spans; for instance, in “the lungs have not been clear”, the **val** is “not ... clear”. This is not a problem as we need to handle noncontiguous spans in other cases as well.

## Coordination:

Q2: How to handle coordination?

A2: The **conj** entity and **combine** relation are used to provide coordination of fields in an event tree. The scope of coordination depends on which entities are being coordinated at different levels of the event tree. Top level **cos** and **dhead** entities are not coordinated.

## Verbs:

Q3: Some verbs such as “appear” in “the lungs appear clear” imply uncertainty or hedging. Should we include them in a field?

A3: While the verbs indeed change the meaning a little bit, for the sake of annotation speed, we are not including verbs in Stage 1. In latter stages, we can find the verb “appear”, the head of “clear”, easily.

## Anaphor resolution:

Q4: A field could be an anaphor or a referencing expression, e.g., “the left one” in “both lungs are otherwise clear; the left one is unchanged”. Should we identify the referent for an anaphor in Stage 1?

A4: The anaphor resolution will be done in Stage 2 for two reasons: first, anaphor is rare because a snippet is often very short; having annotators to worry about anaphor resolution could slow them down. Second, the referent might be outside the snippets, which annotators do not have access to in Stage 1.

## References

- P. Klassen, F. Xia, L. Vanderwende, M. Yetisgen. Annotating Clinical Events in Text Snippets for Phenotype Detection. To Appear in *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland. May, 2014.
- M. Tepper, H.L. Evans, F. Xia, M. Yetisgen-Yildiz. Modeling Annotator Rationales with Application to Pneumonia Classification. *Proceedings of Expanding the Boundaries of Health Informatics Using AI Workshop of AAAI'2013*, 2013.