Contextualized Medication Event Extraction and Cue Detection with Levitated Markers

Annotation guidelines for the evaluation of cue detection

## Running the annotation tool

You will have received a zip file titled cue\_evaluation\_materials.zip containing the following files:

* maskviewer: This is the annotation tool. It is a linux executable
* maskviewer.py: This is the Python source for the executable, meant as a backup if the executable does not work for you.
* annotation\_spec.json: This file defines the annotation task for the annotation tool.
* cue\_predictions/{Action,Actor,Certainty,Negation,Temporality}.json: Files containing the cue predictions for you to evaluate. You will receive a subset of the files listed. Each file contains 50 randomly selected cue predictions for you to evaluate.

Run the annotation tool from the Linux command line

./maskviewer --annotation-spec annotation\_spec.json cue\_predictions/${CONTEXT}.json

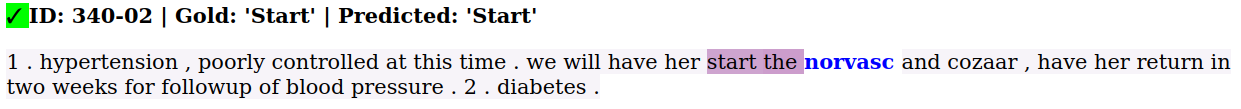
Where ${CONTEXT} is one of the context dimensions in your cue\_predictions/ directory.

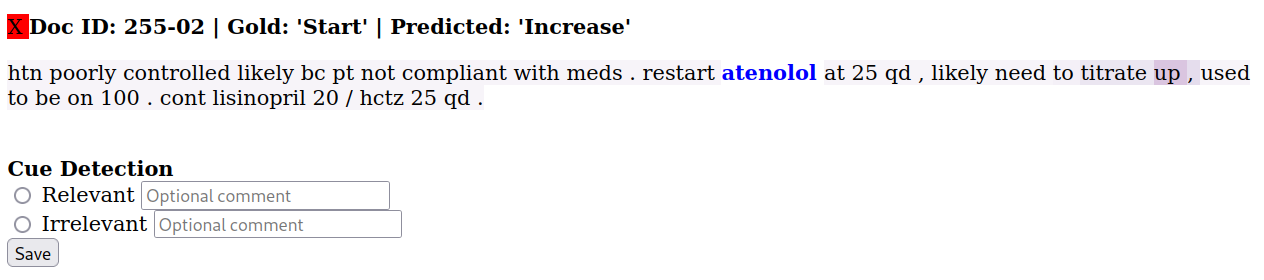
This will load 50 random examples from the predictions\_file and save your annotations to ${CONTEXT}\_annotations.json.

Once the tool is running (you should see \* Running on http://127.0.0.1:5000/ in the output), open your browser and navigate to localhost:5000 to begin annotating.

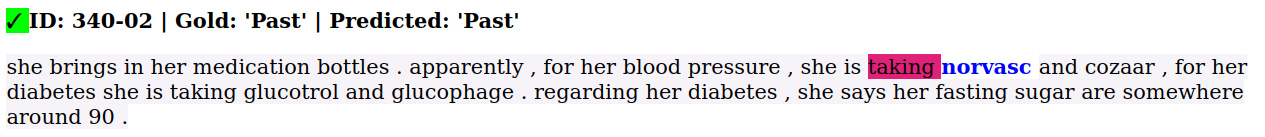
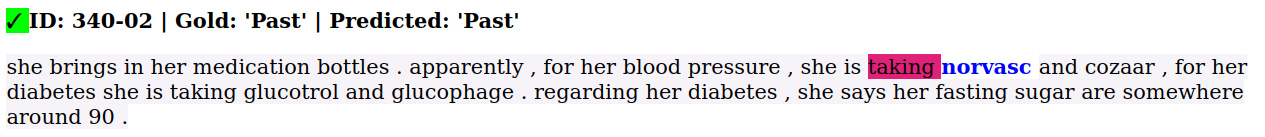
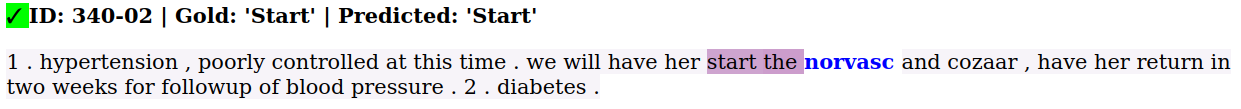
## The annotation task

Upon loading the tool in your browser, you will see something like the following example from the Action dimension.





In this example, you are given the following information:

* **Gold: ‘Start’** means that the gold-standard label for this example is ‘Start’.
* **Predicted: ‘Start’** means that the model predicted the label ‘Start’ for this example.
* The green check  means that the model prediction was correct (i.e., Gold = Predicted)
* The bold blue text is the medication event. In this case,
* The cues detected by the model are highlighted in the text. In this case,

Below that, you have the annotation interface (under the **Cue Detection** heading). You are to mark the cue(s) detected as “Relevant” or “Irrelevant” using the radio buttons.

Determining whether the cues are “Relevant” or “Irrelevant” requires you to answer two questions.

1. Does the highlighted text scope over the medication entity?
2. Is the highlighted text a meaningful cue for the **predicted** label (not the gold label!). That is, given only the highlighted text and the medication event text (in blue), is the predicted label a reasonable one?

If your answer to either 1 or 2 is “no”, the cue is “Irrelevant”, otherwise, the cue is “Relevant”. Record this using the radio buttons. You may also enter an optional comment for each annotation. For example, “The cue is relevant but another non-highlighted word would have been more relevant”.

In the example above, “start the” does scope over the “norvasc” entity (i.e., it belongs to the same verb phrase as norvasc) and is a meaningful cue (i.e., the meaning of “start” in this phrase gives a good clue regarding the predicted label for the Action task), so the cue is “Relevant”.

**Once you’ve selected your answer using the radio buttons, click “Save”, to save your annotation to the output file.**

**You may change your answer any number of times simply by selecting the other radio button and clicking “Save” again.**

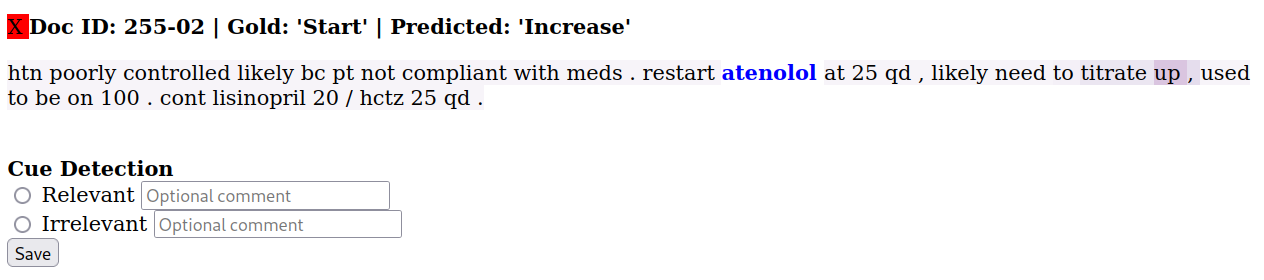
## Notes

You may encounter clinical shorthand you are unfamiliar with. Please do your best to figure out the meaning using Google or other resources. If you do not feel confident in the meaning of a cue, please try your best to make an annotation and record that you’re uncertain in the comment box.

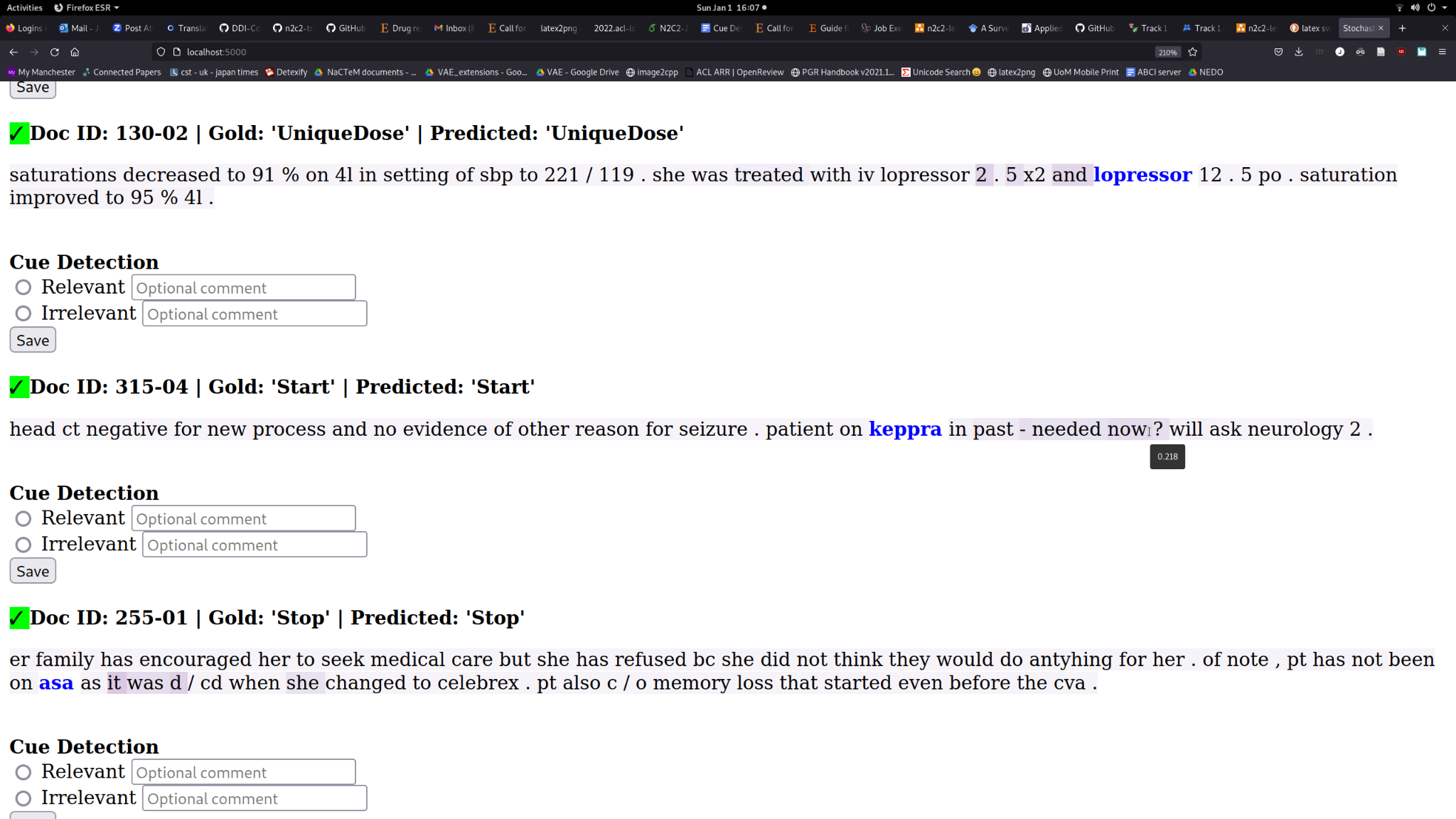
For example, in the following “d/cd” means “discontinued”. As such, it is a “Relevant” cue.

## 

You will likely encounter incorrect model predictions. As stated in the guidelines above, you should **make your decision based on the predicted label only**. For example, the following cue would still be “Relevant”, even though the model prediction is incorrect and the cue for the gold label “Start” should be “restart”.



You may also encounter examples where the cue highlights are quite faint. The strength of the highlight color is proportional to the attention weight given that token. Keep in mind that these weights are probabilities that sum to 1. By hovering your mouse pointer over a token, a little box will appear with the attention weight. For example,



## Description of context dimensions

The table below describes each context dimension and its set of associated labels.

| **Context Dimension** | **Definition** | **Labels** |
| --- | --- | --- |
| Action | What medication change is being discussed. | Start, Stop, Increase, Decrease, OtherChange, UniqueDose |
| Actor | Who initiated the change. | Physician, Patient |
| Certainty | How likely the change is to occur. | Certain, Hypothetical, Conditional |
| Negation | Whether the change is negated. | Negated, NotNegated |
| Temporality | When the change is said to occur. | Past, Present, Future |

For the Actor dimension, it is often the case that the cue does not occur in the immediate context, and is either inferred (often when the label is “Physician”) or determined by the wider, document-level context. Thus, keep in mind that many cues for the Actor dimension may be “Irrelevant”.

For the Certainty and Negation dimensions, there is no “true” cue for “Certain” and “NotNegated” examples, respectively. As such, these labels are excluded from the evaluation.