Contextualized Medication Event Extraction

National NLP Clinical Challenges (n2c2) 2022 Track 1

This document details the expected format for submission for Track 1 and the evaluation scheme. The document is organized as follows. We describe
(1) the annotation format,
(2) the 3 subtasks, provided training data, and evaluation script, and
(3) the test data releases, submission formats for each release, and primary evaluation metrics.

The annotation format

The submission should be in brat standoff format (.ann). The participants need to submit only the .ann files. Please refer to the official documentation for the format details – https://brat.nlplab.org/standoff.html. Please note that the Negation is an explicit attribute with values ('NotNegated', 'Negated') instead of a binary BRAT attribute.

The 3 subtasks, training data, and evaluation script

1. **Medication Extraction [NER]:** Extract all medication mentions in clinical notes. This is a named-entity recognition task.

   Example training data:

   T1<TAB>Disposition<SPACE>1209<SPACE>1214<TAB>Coreg

   where "T1" is the unique ID, "Disposition" is the label for the medication mention "Coreg" with start-offset "1209" and end-offset "1214".
   For this subtask, the evaluation script only considers the offset pair ("1209" & "1214"). Any text (e.g. "Disposition", "Drug") can be used as a label placeholder. The medication text mentioned ("Coreg") is optional and not evaluated.
   Note that the data may contain rows with the same offset pairs but different labels. As labels are not considered for this subtask, only unique offset pairs are counted. That is, multiple annotations sharing the same offsets will only be counted once.

   Evaluation:
   The evaluation script provides strict and lenient evaluation for this subtask. A system annotation is considered a match to a gold annotation if,
   a. Strict Matching: the two offset pairs match exactly
   b. Lenient Matching: there is any overlap between the two offset pairs. Multiple system annotations that overlap with a single gold annotation are counted as one true positive and others are counted as false positives.

Example output from the evaluation script:

```
********************** Medication Extraction ***********************
                    ------- strict -------    ------ lenient -------
                    Prec.   Rec.    F(b=1)    Prec.   Rec.    F(b=1)
                Drug  1.0000  1.0000  1.0000    1.0000  1.0000  1.0000
```

2. **Event Classification [Event]:** Classify medication mentions in clinical notes as either: Disposition (medication change discussed), NoDisposition (no change discussed), or Undetermined (need more information).

   Example training data:
   T1<TAB>Disposition<SPACE>1209<SPACE>1214<TAB>Coreg
   E1<TAB>Disposition:T1

   For this subtask, the evaluation script considers the label ("Disposition" from event E1), as well as the offsets ("1209" & "1214"). The medication text mentioned ("Coreg") is optional and not evaluated.
   Note that the data may contain multiple rows with the same labels and the same offset pairs. For this subtask, only unique Event annotations are counted (i.e. annotations with the same label and the same offset pair will only be counted once).

   Evaluation:
   The evaluation script provides P/R/F_1 (strict & lenient) for the 3 event types – Disposition, NoDisposition & Undetermined. For lenient evaluation, multiple system annotations that overlap with a single gold annotation and have the same label as the gold annotation are counted as one true positive and others are counted as false positives. Event labels are case-sensitive.
   Example output from the evaluation script:

```
************************ Event Classification ************************
                      ------- strict -------    ------ lenient -------
                      Prec.   Rec.    F(b=1)    Prec.   Rec.    F(b=1)
          Disposition  0.6667  0.6667  0.6667    1.0000  1.0000  1.0000
        Nodisposition  1.0000  1.0000  1.0000    1.0000  1.0000  1.0000
         Undetermined  1.0000  1.0000  1.0000    1.0000  1.0000  1.0000
                      ------------------------------------------------
      Overall (micro)  0.8333  0.8333  0.8333    1.0000  1.0000  1.0000
      Overall (macro)  0.8889  0.8889  0.8889    1.0000  1.0000  1.0000
```

3. **Context Classification [Context]**: Classify the contextual information for Disposition events along 5 orthogonal dimensions: Action (e.g. Start, Stop), Negation (e.g. Negated), Temporality (e.g. Past, Present), Certainty (e.g. Hypothetical, Conditional), and Actor (e.g. Patient, Physician). For the full set of values associated with each context dimension, please refer to the CMED dataset paper – "Toward Understanding Clinical Context of Medication Change Events in Clinical Narratives (attached CMED.pdf)".

Example training data:

T1<TAB>Disposition<SPACE>1209<SPACE>1214<TAB>Coreg
E1<TAB>Disposition:T1
A1<TAB>Action<SPACE>E1<SPACE>Start
A2<TAB>Temporality<SPACE>E1<SPACE>Past
A3<TAB>Certainty<SPACE>E1<SPACE>Certain
A4<TAB>Actor<SPACE>E1<SPACE> Patient
A5<TAB>Negation<SPACE>E1<SPACE>NotNegated

Please note that there is a 1:1 mapping between entities(T) and events(E). If there are two events associated with a single drug, two entities and two events will be present.

Example training data:

T1<TAB>Disposition<SPACE>1209<SPACE>1214<TAB>Coreg
E1<TAB>Disposition:T1
A1<TAB>Action<SPACE>E1<SPACE>Start
A2<TAB>Temporality<SPACE>E1<SPACE>Past
A3<TAB>Certainty<SPACE>E1<SPACE>Certain
A4<TAB>Actor<SPACE>E1<SPACE> Patient
A5<TAB>Negation<SPACE>E1<SPACE>NotNegated

T2<TAB>Disposition<SPACE>1209<SPACE>1214<TAB>Coreg
E2<TAB>Disposition:T2
A6<TAB>Action<SPACE>E2<SPACE>Stop
A7<TAB>Temporality<SPACE>E2<SPACE>Past
A8<TAB>Certainty<SPACE>E2<SPACE>Certain
A9<TAB>Actor<SPACE>E2<SPACE> Patient
A10<TAB>Negation<SPACE>E2<SPACE>NotNegated

Evaluation:

The evaluation script provides P/R/F1 (strict & lenient) for each of the 5 context dimensions independently. The script also evaluates all 5-dimensional values for a medication combined together as “Combined” (e.g., “Start” + “Past” + “Certain” + “Physician” + “NotNegated”). Attribute values are case-sensitive.

Example output from the evaluation script:

```
********************** Context Classification **********************
                     ------- strict -------    ------ lenient -------
                     Prec.   Rec.    F(b=1)    Prec.   Rec.    F(b=1)
            Action   0.7500  0.7500  0.7500    0.7500  0.7500  0.7500
      Temporality    0.7500  0.7500  0.7500    0.7500  0.7500  0.7500
        Certainty    0.7500  0.7500  0.7500    0.7500  0.7500  0.7500
            Actor    0.7500  0.7500  0.7500    0.7500  0.7500  0.7500
         Negation    0.7500  0.7500  0.7500    0.7500  0.7500  0.7500
                     ------------------------------------------------
  Overall (micro)    0.7500  0.7500  0.7500    0.7500  0.7500  0.7500
  Overall (macro)    0.7500  0.7500  0.7500    0.7500  0.7500  0.7500


                     ------- strict -------    ------ lenient -------
                     Prec.   Rec.    F(b=1)    Prec.   Rec.    F(b=1)
          Combined   0.2500  0.2500  0.2500    0.2500  0.2500  0.2500
```

The test data releases, submission formats, and primary evaluation metrics

Test data for this track will be released in 3 stages. Each team can submit up to 3 runs for each release. During submission time, for each submission, teams need to specify which subtask(s) the submission is for. All submissions should be in brat standoff format, similar to the gold annotations. The release stages are detailed below:

1. **Release 1:**
   This release will only have .txt files containing the clinical note text.
   Participants may submit .ann files containing NER, NER+Event, or NER+Event+Context annotations.

   Sample submission 1 (.ann file containing only NER annotations):
   T1<TAB>Drug<SPACE>1209<SPACE>1214<TAB>Coreg

   Sample submission 2 (.ann file containing NER+Event annotations):
   T1<TAB>Drug<SPACE>1209<SPACE>1214<TAB>Coreg
   E1<TAB>Disposition:T1

   Sample submission 3 (.ann file containing NER+Event+Context annotations):
   T1<TAB>Drug<SPACE>1209<SPACE>1214<TAB>Coreg
   E1<TAB>Disposition:T1
   A1<TAB>Action<SPACE>E1<SPACE>Start
   A2<TAB>Temporality<SPACE>E1<SPACE>Past
   A3<TAB>Certainty<SPACE>E1<SPACE>Certain
   A4<TAB>Actor<SPACE>E1<SPACE>Patient
   A5<TAB>Negation<SPACE>E1<SPACE>NotNegated

   For the official results, the primary evaluation metrics for this release are:
   - NER: Strict F_1-score and Lenient F_1-score

- NER+Event: Lenient micro F_1-score & Lenient macro F_1-score
- NER+Event+Context(E2E): Combined lenient F_1-score

2. **Release 2:**
   This release will have .txt files containing the clinical note text and .ann files containing the gold NER annotations labeled as "Drug".

   Sample provided .ann file:
   T1<TAB>Drug<SPACE>1209<SPACE>1214<TAB>Coreg

   Participants may submit .ann files containing Event or Event+Context annotations.

   Sample submission 1 (.ann file containing only Event annotations):
   T1<TAB>Drug<SPACE>1209<SPACE>1214<TAB>Coreg
   E1<TAB>Disposition:T1

   Sample submission 2 (.ann file containing Event+Context annotations):
   T1<TAB>Drug<SPACE>1209<SPACE>1214<TAB>Coreg
   E1<TAB>Disposition:T1
   A1<TAB>Action<SPACE>E1<SPACE>Start
   A2<TAB>Temporality<SPACE>E1<SPACE>Past
   A3<TAB>Certainty<SPACE>E1<SPACE>Certain
   A4<TAB>Actor<SPACE>E1<SPACE>Patient
   A5<TAB>Negation<SPACE>E1<SPACE>NotNegated

   For the official results, the primary evaluation metrics for this release are:
   - Event: Lenient micro F_1-score & Lenient macro F_1-score
   - Event+Context: Combined lenient F_1-score

3. **Release 3:**
   This release will have .txt files containing the clinical note text and .ann files containing the NER+Event annotations.

   Sample provided .ann file:

   T1<TAB>Disposition<SPACE>1209<SPACE>1214<TAB>Coreg
   E1<TAB>Disposition:T1
   T2<TAB>Undetermined<SPACE>1707<SPACE>1712<TAB>Lasix

   Participants may submit .ann files containing Context annotations. Note that the provided .ann files will contain all events (i.e., Disposition, NoDisposition, and Undetermined). However, participants only need to submit Context annotations for Disposition events (i.e., can ignore NoDisposition and Undetermined annotations).

Sample submission (.ann file containing 2 Context annotations for the same span Coreg):
T1<TAB>Disposition<SPACE>1209<SPACE>1214<TAB>Coreg
E1<TAB>Disposition:T1
T2<TAB>Undetermined<SPACE>1707<SPACE>1712<TAB>Lasix

A1<TAB>Action<SPACE>E1<SPACE>Start
A2<TAB>Temporality<SPACE>E1<SPACE>Past
A3<TAB>Certainty<SPACE>E1<SPACE>Certain
A4<TAB>Actor<SPACE>E1<SPACE>Patient
A5<TAB>Negation<SPACE>E1<SPACE>NotNegated

T2<TAB>Disposition<SPACE>1209<SPACE>1214<TAB>Coreg
E2<TAB>Disposition:T2
A6<TAB>Action<SPACE>E2<SPACE>Stop
A7<TAB>Temporality<SPACE>E2<SPACE>Past
A8<TAB>Certainty<SPACE>E2<SPACE>Certain
A9<TAB>Actor<SPACE>E2<SPACE>Patient
A10<TAB>Negation<SPACE>E2<SPACE>NotNegated

For the official results, the primary evaluation metric for this release is the Combined lenient F_1-score.