ETUDE for Easy and Efficient NLP Application Evaluation

Stéphane M. Meystre, MD, PhD[1,2], Paul M. Heider, PhD[1],
Jean-Karlo Accetta, MSc[2]

[1] Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC

[2] Clinacuity, Inc., Charleston, SC

OHDSI NLP Working Group Webinar
May 8, 2019

Agenda

1. The Problem Space
2. System Overview

Introduction

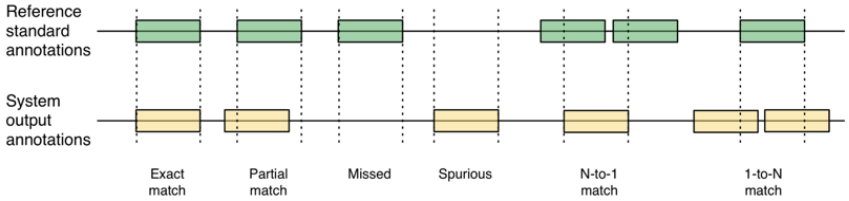
NLP development and evaluation requires tools to compare annotations, analyze differences, and compute accuracy metrics.

Only few tools already exist for this purpose such as:

- Evaluation Workbench (Lee Christensen, Utah; Java development compatible with Knowtator and eHost, computes accuracy metrics and displays annotations)
- i2b2 challenge evaluation scripts (Java applications or Python scripts, without GUI, computes accuracy metrics)
- WebAnno (Web and Java application, computes agreement between annotation sets)
- Brat (Web application, displays matching and mismatching annotations)

Introduction

None of these tools allows consuming the output of any NLP system, comparing it with reference annotations in any format (both with configuration), setting comparison rules, counting matches and computing accuracy metrics, visually displaying matching and erroneous annotations, changing the displayed information for analysis on the fly, exporting evaluation results, etc.

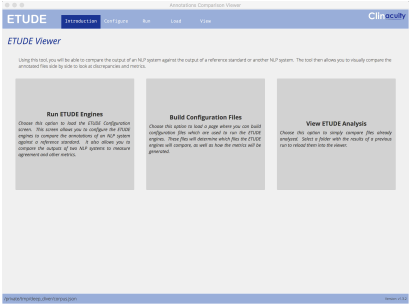


Functional Requirements

- Import of annotations (system output and reference standard) in multiple configurable formats
- Drag-and-drop creation of evaluation configuration files (i.e., specification of annotation categories and attributes to match and compare)
- Customizable annotation comparison settings (base includes exact, partial, and fully-contained matches)
- Count and comparison of annotations, computing descriptive statistics, a confusion matrix of matches and mismatches, and computing accuracy metrics (recall, precision, F-measure)
- Show side-by-side reference and system annotations in the original document context with easy exploration of matches and errors
- Export of evaluation results in various configurable formats
- Comparison based on annotation attributes
- Comparison based on normalized annotations (e.g., matched with UMLS concepts) with hierarchical matches (i.e., including select parent/child or related concepts)
- Comparison/analysis of evaluations history

New NLP Evaluation Tool

Evaluation **T**ool for **U**nstructured **D**ata **E**xtractions **(ETUDE)**



Viewer (Java)

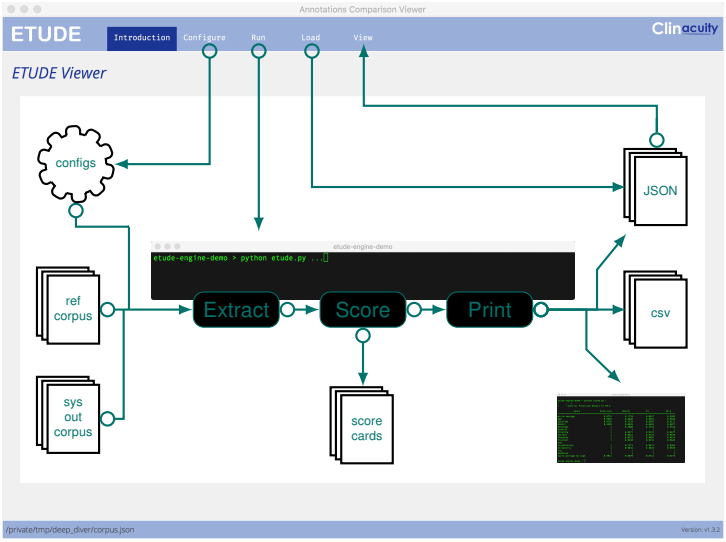
`https://github.com/Clinacuity/etude-viewer`

+

Engine (Python)

`https://github.com/MUSC-TBIC/etude-engine`

Clinacuity



```
etude-engine-demo > python etude.py \
>     --reference-input tests/data/i2b2_2016_track-1_reference \
>     --test-input tests/data/i2b2_2016_track-1_test

exact  TP  FP  TN  FN
micro-average  374.0  8.0  0.0  108.0
```

```
etude-engine-demo > python etude.py \
>     --reference-input tests/data/i2b2_2016_track-1_reference \
>     --test-input tests/data/i2b2_2016_track-1_test \
>     --progressbar-output none \
>     --pretty-print

            exact             TP      FP      TN      FN
=============================================================
micro-average               374       8       0     108
```

```
etude-engine-demo > python etude.py \
>     ... \
>     --by-type
```

exact	TP	FP	TN	FN
micro-average	374	8	0	108
Age	63	2	0	29
DateTime	91	2	0	33
HCUnit	61	4	0	15
OtherGeo	1	0	0	4
OtherID	7	0	0	0
OtherOrg	18	0	0	3
...				
macro-average by type	374	8	0	108

```
etude-engine-demo > python etude.py \
>     ... \
>     --metrics Precision Recall F1 F0.5
```

exact	Precision	Recall	F1	F0.5
micro-average	0.9791	0.7759	0.8657	0.9303
Age	0.9692	0.6848	0.8025	0.8949
DateTime	0.9785	0.7339	0.8387	0.9173
HCUnit	0.9385	0.8026	0.8652	0.9077
...				
SSN				
...				
macro-average by type	0.9912	0.8070	0.8712	0.9274

```
etude-engine-demo > python etude.py \
>      ... \
>      --score-value "Age|DateTime|HCUnit"
```

exact	Precision	Recall	F1	F0.5
micro-average	0.9641	0.7363	0.8350	0.9079
Age	0.9692	0.6848	0.8025	0.8949
DateTime	0.9785	0.7339	0.8387	0.9173
HCUnit	0.9385	0.8026	0.8652	0.9077
macro-average by type	0.9621	0.7404	0.8355	0.9067

```
etude-engine-demo > python etude.py \
>     ... \
>     --no-metrics \
>     --print-confusion-matrix

exact  *FP*  Age  DateTime  HCUnit
*FN*    27  30  12
Age    63  1  1
DateTime      91  2
HCUnit    2  2  61
```

```
etude-engine-demo > python etude.py \
>     ... \
>     --print-confusion-matrix \
>     --delim " | "

exact | *FP* | Age | DateTime | HCUnit
*FN* |  | 27 | 30 | 12
Age |  | 63 | 1 | 1
DateTime |  |  | 91 | 2
HCUnit |  | 2 | 2 | 61
```

```
etude-engine-demo > python etude.py \
>     ... \
>     --print-confusion-matrix \
>     --delim " | "

...

exact | TP | FP | TN | FN
micro-average | 215.0 | 8.0 | 0.0 | 77.0
Age | 63.0 | 2.0 | 0.0 | 29.0
DateTime | 91.0 | 2.0 | 0.0 | 33.0
HCUnit | 61.0 | 4.0 | 0.0 | 15.0
macro-average by type | 215.0 | 8.0 | 0.0 | 77.0
```

```
etude-engine-demo > python etude.py \
>     --reference-input tests/data/offset_matching/reference \
>     --reference-config config/webanno_phi_xmi.conf \
>     --test-input tests/data/offset_matching/system_out \
>     --test-config config/webanno_phi_xmi.conf \
>     ... \
>     --file-suffix ".xmi" \
>     --fuzzy-match-flag exact partial fully-contained
```

exact	TP	FP	FN
micro-average	14	17	16
Age	11	4	4
DateTime	3	13	12
macro-average by type	14	17	16

```
etude-engine-demo > python etude.py \
>      ...
>      --fuzzy-match-flag exact partial fully-contained

             exact                    TP         FP         FN
================================================================
micro-average                        14         17         16
...
             partial                  TP         FP         FN
================================================================
micro-average                        22          9          8
Age                                  11          4          4
DateTime                             11          5          4
macro-average by type                22          9          8
...
```

```
           exact                    TP          FP          FN
=============================================================
micro-average                      14          17          16
...
          partial                   TP          FP          FN
=============================================================
micro-average                      22           9           8
...
       fully-contained              TP          FP          FN
=============================================================
micro-average                      17          14          13
Age                                11           4           4
DateTime                            6          10           9
macro-average by type              17          14          13
```

```
etude-engine-demo > python etude.py \
>     ....
>     --collapse-all-patterns

            exact                    TP         FP         FN
============================================================
micro-average                       18         13         12

           partial                   TP         FP         FN
============================================================
micro-average                       28          3          2

       fully-contained               TP         FP         FN
============================================================
micro-average                       23          8          7
```

```
etude-engine-demo > python etude.py \
>      ....
>      --by-type
```

exact	TP	FP	TN	FN
micro-average	203	13	0	90
Allergen	7	1	0	2
Problem	196	12	0	88
macro-average by type	203	13	0	90

```
etude-engine-demo > python etude.py \
>     ....
>     --by-type \
>     --by-type-and-attribute \
>     --score-attributes
```

exact	TP	FP	TN	FN
micro-average	203	13	0	90
...				
Problem	196	12	0	88
Problem x conditional	1	5	188	2
Problem x negated	28	4	163	1
...				
macro-average by type	203	13	0	90

```
etude-engine-demo > python etude.py \
>      ....
>      --by-attribute \
>      --score-attributes
```

exact	TP	FP	TN	FN
...				
conditional	1	5	195	2
generic	0	16	187	0
historical	17	1	138	47
negated	28	4	170	1
not_patient	11	0	192	0
uncertain	6	5	178	14
macro-average by pivot	63	31	1060	64

```
<custom:Problems
    CUI="C0004096"
    Problem="asthma"
    begin="658" end="664"
    conditional="false" generic="false" historical="false"
    negated="false" not_patient="false" sofa="12" uncertain="fals
    xmi:id="9680"
/>
<conceptMapper:DictTerm
    xmi:id="9318" sofa="1"
    begin="1452" end="1466"
    PreferredTerm="Abdominal␣Pain,␣CTCAE␣5"
    CUI="C4554323" TUI="T033"
    TermType="Problem"
    enclosingSpan="8" matchedTokens="2231␣2241"
/>
```

```
etude-engine-demo > python etude.py \
>      ....
>      --score-normalization "CUI"

            exact             TP          FP          TN
============================================================
micro-average                108         831         60

          exact_CUI           TP          FP          TN
============================================================
micro-average                 97          11          11

>      --score-normalization "CUI/umlsCode"
...
>      --score-normalization "CUI/ID"
```

```
etude-engine-demo > python etude.py \
>      ....
>      --score-normalization "CUI/PreferredTerm" \
>      --normalization-file cui2pt.csv

etude-engine-demo > python etude.py \
>      ....
>      --score-normalization "CUI/Term" \
>      --normalization-file cui_and_synonyms.csv
```

```
etude-engine-demo > python etude.py \
>       ...
>       --print-custom "2018 n2c2 track 1"
```

```
****************************************** TRACK 1 ******************************************
                    ------------ met -------------     ------ not met -------     -- overall ---
                    Prec.   Rec.    Speci.  F(b=1)    Prec.   Rec.    F(b=1)     F(b=1)  AUC
          Abdominal 0.9697  0.4156  0.9920  0.5818    0.7337  0.9920  0.8435     0.7127  0.7038
       Advanced-cad 0.7500  0.8640  0.5325  0.8030    0.7069  0.5325  0.6074     0.7052  0.6982
    Alcohol-abuse   0.0000  0.0000  0.9846  0.0000    0.9648  0.9846  0.9746     0.4873  0.4923
       Asp-for-mi   0.8870  0.9632  0.4872  0.9235    0.7600  0.4872  0.5938     0.7586  0.7252
       Creatinine   0.8214  0.8415  0.8750  0.8313    0.8898  0.8750  0.8824     0.8568  0.8582
    Dietsupp-2mos   0.9302  0.3774  0.9688  0.5369    0.5849  0.9688  0.7294     0.6332  0.6731
       Drug-abuse   0.5000  0.3000  0.9844  0.3750    0.9643  0.9844  0.9742     0.6746  0.6422
          English   0.9552  1.0000  0.1000  0.9771    1.0000  0.1000  0.1818     0.5795  0.5500
            Hba1c   0.8000  0.7164  0.9111  0.7559    0.8662  0.9111  0.8881     0.8220  0.8138
        Keto-1yr    0.0000  0.0000  1.0000  0.0000    1.0000  1.0000  1.0000     0.5000  0.5000
   Major-diabetes   1.0000  0.3982  1.0000  0.5696    0.5669  1.0000  0.7236     0.6466  0.6991
  Makes-decisions   0.9648  0.9897  0.1250  0.9771    0.3333  0.1250  0.1818     0.5795  0.5573
          Mi-6mos   0.8000  0.2222  0.9946  0.3478    0.9289  0.9946  0.9606     0.6542  0.6084
                    ----------------------------      ----------------------     --------------
  Overall (micro)   0.8900  0.7712  0.9253  0.8264    0.8376  0.9253  0.8793     0.8528  0.8483
  Overall (macro)   0.7214  0.5452  0.7658  0.5907    0.7923  0.7658  0.7339     0.6623  0.6555

202 files found
```

```
etude-engine-demo > python etude.py \
>       ...
>       --print-custom "2018 n2c2 track 1" \
>       --file-suffix "3.xml"

****************************************** TRACK 1 ******************************************
                    ------------ met -------------     ------ not met -------     -- overall ---
                    Prec.   Rec.    Speci.  F(b=1)     Prec.   Rec.    F(b=1)     F(b=1)  AUC
          Abdominal 1.0000  0.3750  1.0000  0.5455     0.6429  1.0000  0.7826     0.6640  0.6875
       Advanced-cad 0.6429  0.9000  0.2857  0.7500     0.6667  0.2857  0.4000     0.5750  0.5929
      Alcohol-abuse 0.0000  0.0000  1.0000  0.0000     1.0000  1.0000  1.0000     0.5000  0.5000
         Asp-for-mi 0.9286  1.0000  0.7500  0.9630     1.0000  0.7500  0.8571     0.9101  0.8750
         Creatinine 0.8333  0.7143  0.9000  0.7692     0.8182  0.9000  0.8571     0.8132  0.8071
      Dietsupp-2mos 1.0000  0.3750  1.0000  0.5455     0.6429  1.0000  0.7826     0.6640  0.6875
         Drug-abuse 0.0000  0.0000  1.0000  0.0000     1.0000  1.0000  1.0000     0.5000  0.5000
            English 1.0000  1.0000  0.0000  1.0000     0.0000  0.0000  0.0000     0.5000  0.5000
              Hba1c 0.6667  0.5000  0.9231  0.5714     0.8571  0.9231  0.8889     0.7302  0.7115
          Keto-1yr  0.0000  0.0000  1.0000  0.0000     1.0000  1.0000  1.0000     0.5000  0.5000
    Major-diabetes  1.0000  0.4444  1.0000  0.6154     0.6154  1.0000  0.7619     0.6886  0.7222
   Makes-decisions  1.0000  0.9412  0.0000  0.9697     0.0000  0.0000  0.0000     0.4848  0.4706
          Mi-6mos   0.0000  0.0000  1.0000  0.0000     1.0000  1.0000  1.0000     0.5000  0.5000
                    ----------------------------     ----------------------     --------------
   Overall (micro)  0.9000  0.7742  0.9375  0.8324     0.8511  0.9375  0.8922     0.8623  0.8558
   Overall (macro)  0.6209  0.4808  0.7584  0.5177     0.7110  0.7584  0.7177     0.6177  0.6196

                                    17 files found
```

This work was supported by:
the SmartState Program (Translational Biomedical Informatics Chair Endowment)
&
the National Cancer Institute (5R42CA180190).

Software:

`https://github.com/MUSC-TBIC/etude-engine`
`https://github.com/MUSC-TBIC/etude-engine-configs`
`https://github.com/Clinacuity/etude-viewer`

Documentation

`https://etude-engine.readthedocs.io`