Annotation Guideline for Rare Disease Identification from Clinical Notes

This is the guideline of annotation for rare disease identification from clinical notes. MIMIC-III discharge summaries and radiology reports were used in this study [1]. The guideline will first explain the two annotation tasks, Mention-to-UMLS annotation and UMLS-to-ORDO annotation, then will introduce the annotation tools, Brat and Excel sheets.

1 Annotation Tasks

1.1 Mention-to-UMLS annotation

If a mention in a section of a clinical note matches to the UMLS concept AND is the **phenotype** of that patient, then mark as True (indicated as "1" in excel or **true_phenotype** in Brat), otherwise False ("0" or **false_phenotype**).

Phenotype means that the disease is the **positive** mention of the patient. Past medical history also counts as a positive mention. The below example is False due to the mention is negative.

> `urinary legionella antigen was also negative (legionella: C0023241 Legionella)`

For hypothetical mentions, consider the likelihood of this rare disease for the patient given the context. If *likely*, then mark as True; if just *somewhat likely*, then mark as False.

Below is a *likely* case of a rare disease considering that the section is in discharge diagnosis with the word "potential", thus mark as True.

> `Discharge Diagnosis: … 3. Potential autoimmune pancreatitis. (autoimmune pancreatitis: C2609129 Autoimmune pancreatitis)`

Below is a case of *somewhat likely*, based on the context "`raises the possibility of`", thus mark as False.

> `The amount and distribution of iron deposition raises the possibility of a genetic iron storage disease (iron storage disease: C0018995 Hemochromatose)`

Below is a case of *unsure* based on the context that this is only an examination, thus mark as False.

> `REASON FOR THIS EXAMINATION: eval for cholecystitis, portal vein thrombosis (portal vein thrombosis: C0155773_Pylethrombosis)`

The annotator should mainly refer to the official UMLS browser, given below. Other sources can only be used if the source below is not enough to make a decision.

> Official UMLS browser https://uts.nlm.nih.gov/uts/umls/home (a user can sign in with a Google account or with institutional account)

To note that the name of the UMLS in the annotation sheet is not always the preferred label. The only identifier to specify a UMLS concept is the CUI (Concept Unique Identifier, the 8-character unique code, e.g. *C0020473*).

1.2 UMLS-to-ORDO annotation

If a UMLS concept and an ORDO concept describe exactly the same rare disease, then mark as True (indicated as "1"), otherwise False ("0").

The annotator should mainly refer to the official UMLS browser and ORDO browser, given below. Other sources can only be used if the two sources are not enough to make a decision.

> Official UMLS browser https://uts.nlm.nih.gov/uts/umls/home (a user can sign in with a Google account or with institutional account)
>
> Official ORDO browser https://www.ebi.ac.uk/ols/ontologies/ordo

To note that the name of the UMLS in the annotation sheet is not always the preferred label. The only identifier to specify a UMLS concept is the CUI (Concept Unique Identifier, the 8-character unique code, e.g. *C0020473*).

For accurate rare disease detection, a narrower ORDO concept should not match to a broader UMLS concept, in this case, the annotation should be False ("0"). A false matching example is below, where Hyperlipidemia is more general than "Rare hyperllipidemia".

> *UMLS/C0020473 Hyperlipidemia* to *Orphanet_181422 Rare hyperlipidemia*

If an ORDO concept has a more specific UMLS concept than the current one, then should be False ("0"). For the example below, the best match to Orphanet_166282 should be UMLS/C0340491 (Familial sick sinus syndrome).

> *UMLS/C0037052 Sick Sinus Syndrome* to *Orphanet_166282 Familial sick sinus syndrome*

2 Annotation Tools and Interfaces

2.1 Brat – for Mention-to-UMLS annotation only

To get access to Brat interface, the user need to send an `ssh` command from their local computer.

```
ssh -N -L <PORT_NUMBER>:localhost:<PORT_NUMBER> -i <YOUR_PRIVATE_KEY>
<YOUR_NAME>@<SERVER_IP_ADDRESS>
```

Then visit http://localhost:<PORT_NUMBER>/brat-v1.3_Crunchy_Frog from the browser.

(run **docker restart rd_ann_<ANNOTATOR_INITIAL_LETTERS>** from the KnowLab server if console shows channel 2: open failed: connect failed: Connection refused)

To login, hover the mouse on the top of the brat interface and click the "login" button on the right. Then login with `username`//`password`.

Please contact us for PORT_NUMBER, YOUR_PRIVATE_KEY, `SERVER_IP_ADDRESS`, ANNOTATOR_INITIAL_LETTERS**,** username, and password and to report any issues.

To annotate with the Brat interface (see Figure 1 below),

(i) double click the green highlighted annotation, and then select one from **true_phenotype, false_phenotype, or unsure** in the "entity attributes" (see bottom-left of the screencast below). You can select unsure only (or also with one in true_phenotype or false_phenotype) if you are not certain about this mention.

(ii) Untick the "**positive**" attribute if the mentioned disease is in a negative context.

(iii) Add notes if necessary.

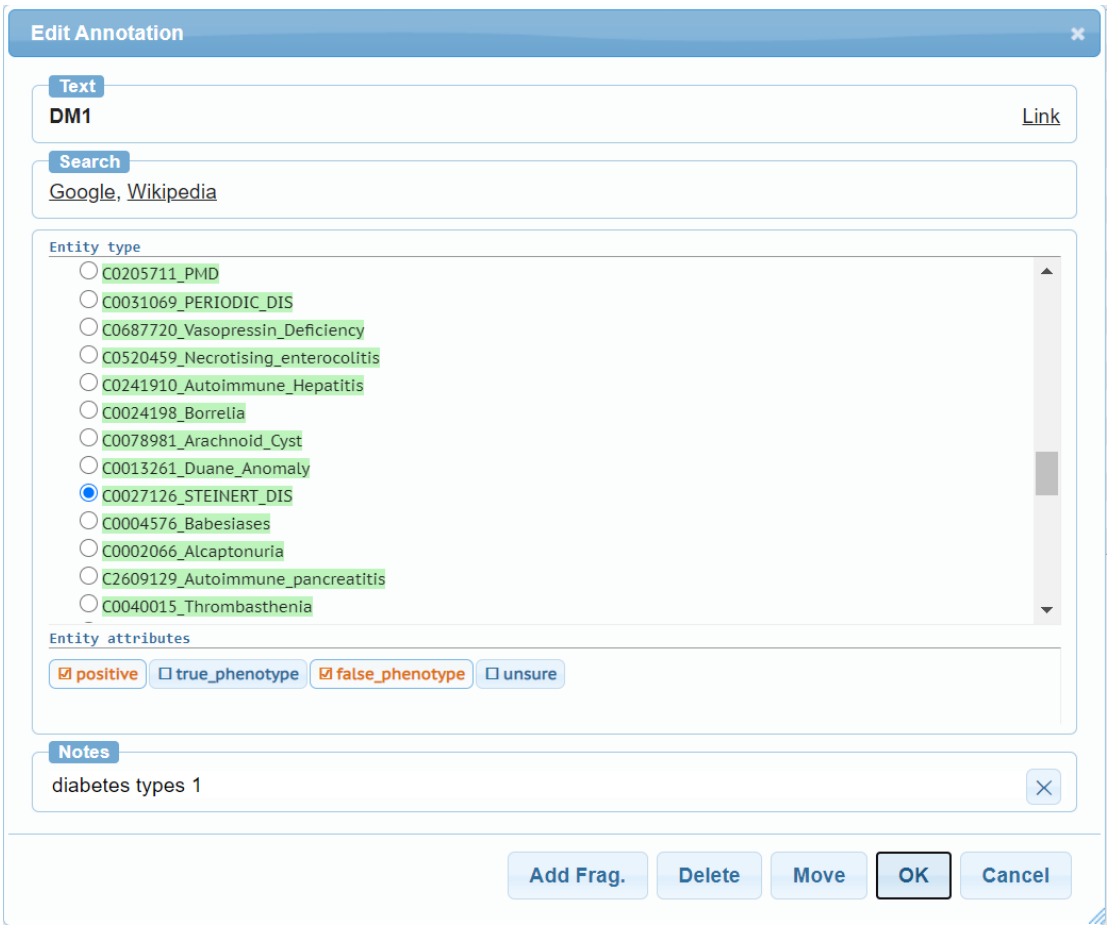

Figure 1. The Brat Annotation Interface

2.2 Excel sheet

In Microsoft Excel, for Mention-to-UMLS annotation, given data from the first four columns (document structure, text, mention, UMLS with description), add the annotation as 1 or 0 to the new, fifth column and may add notes if necessary.

The fields of the first four columns are below.

document structure	**Text**	**mention**	**UMLS with desc**

For UMLS-to-ORDO annotation, given a pair of UMLS and ORDO concepts (each with a description) in two columns, annotator 1 or 0 in the column on the right.

UMLS with desc	**ORDO with desc**

[1] H. Dong, V. Suárez-Paniagua, H. Zhang, M. Wang, E. Whitfield, and H. Wu, 'Rare Disease Identification from Clinical Notes with Ontologies and Weak Supervision', in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, Nov. 2021, pp. 2294–2298. arXiv: 2105.01995

Last Edit: 17 Mar 2022 by Hang Dong