Annotation Guidelines

Annotation Training
Healthcare

2023

Content

The goal of annotations is to train Natural Language Processing (NLP) models to extract relevant information from different kinds of documents.

Annotation Guidelines (AG) define the entities to be extracted and the way they have to be extracted, with the aim of assuring consistency between annotators.

Extraction of entities, relations and assertions will be explained in these AG.

Taxonomy

1) Entities
    a) Age
    b) Gender
    c) Employment
    d) Date
    e) Disease_Syndrome_Disorder
    f) Modifier
    g) VS_Finding
    h) Test_Result
    i) Drug_Ingredient
2) Assertions
    a) Present
    b) Absent
    c) Possible
    d) Family
3) Relations
    a) Is_diagnosis_date_of
    b) Is_modifier_of
    c) Is_cause_of
    d) Is_result_of
4) Text Classification
    a) Gender
    b) Type of text

Entity Labels

For each kind of entity, these AG provide a definition, specific extraction rules, examples, relations with other variables and assertion labels. Examples only include the annotations that are relevant for the kind of entity that is being explained. The following color reference is used:

- Green: Correct extraction.
- Grey: Incorrect extraction.

Every mention of a relevant entity must be extracted, even if it is repeated or negated. Punctuation marks should not be extracted unless stated otherwise. Extractions should neither start nor finish with a blank space.

Age

Definition: All mention of ages, past or present, related to the patient or with anybody else.
Extraction rules: Numbers should be extracted along with expressions such as “years old” or “age of”. Ages expressed in terms of decades should also be extracted.
Examples:

a) The patient presents as a 66-year-old Caucasian female in stable health.
b) The patient was diagnosed in his 50s.
c) John is a 65 years old male.
d) He was diagnosed with obesity at the age of 25.

Assertion labels : None.
Relations : None.

Gender

Definition: Gender-specific nouns, excluding family members.

Extraction rules: The gender label should be used for gender specific nouns, except for family members such as “father” or “sister”. Pronouns (such as “his” or “her” should not be extracted.

Examples:

a) The patient is a 66-year-old lady in stable health.
b) She was treated with tamoxifen.
c) He was intrigues by her alcohol consumption habit.
d) His father was diagnosed with dementia at 55 years old.

Assertion labels : None.
Relations : None.

Employment

Definition: Mentions of jobs or occupations included in the text.

Extraction rules: Extract terms that are related to any specific jobs or employment,whether related to the patient or not. Do not extract words such as "works", "working" or "employed".

Examples:

a) She is an office manager for a gravel company.
b) She will also see a nutritionist and a social worker.
c) He works as a financial officer.

Assertions: None.

Relations: None.

Date

Definition: Mention of an exact date, in any format, including day number, month and/or year.

Extraction rules: Months can be expressed as numbers or in words. All the parts of the date should be included in the same extraction (i.e., day, month and year that refer to the same date should not be split if they appear together in the text). Do not extract relative dates i.e., today, tomorrow, yesterday, etc.

Examples:

a) The patient was diagnosed with diabetes in 1998.

b) On 05/04/2015, the patient started chemotherapy.

c) Yesterday, the patient started metformin.

d) It should be administered in the morning.

Assertions: None.
Relations: is_diagnosis_date_of.

Disease_Syndrome_Disorder

Definition: Extract all the diseases, syndromes and any relevant condition mentioned in the document.

Extraction rules: Extract all mentions of medical conditions and diseases, including those related to the patient or to a family member. Do not include in the extraction modifiers such as “chronic”, “mild” or “severe” (this kind of words should be extracted using the label Modifier).

Examples:

a) The patient has Alzheimer diagnosed back in 2012.
b) He was diagnosed with colon cancer.
c) A diagnosis of chronic kidney disease was established in the past.
d) He was diagnosed twice with chronic depression.

Assertions: Present (by default), Absent, Possible, Family.

Relations: Is_diagnosis_date_of, Is_modifier_of, Is_cause_of.

Modifier

Definition: Terms that modify the medical problem.

Extraction rules: Extract words that indicate severity (such as “mild” or “severe”), duration (such as “chronic” or “acute”) or any other feature of the entities.

Examples:

a) He has been experiencing chronic back pain for five years.
b) Patient with history of recurrent angina.

Assertions: None.

Relations: Is_modifier_of.

VS_Finding

Definition: Vital sign (VS) finding - Names of any vital sign present in the text (heart rate, blood pressure,respiratory rate, temperature and oxygen saturation).

Extraction rules: It only include the names of the vital signs, such as “Blood pressure” (or “BP”), “Pulse”, “Respiratory Rate” (or “RR”), “Temperature”, etc. Do not extract vital sign measurements.

Examples:

a) VITAL SIGNS: Stable blood pressure and respiratory rate.
b) His heart rate is high.

Assertions: None.

Relations: Is_result_of.

Test_Result

Definition: Includes the Vital Signs test Measurement.

Extraction rules: Extract together the numerical values and the units of measurement (such as “mmHg” or “bpm”) of the vital signs. Do not extract the name of the vital signs, as they are to be extracted under VS_Finding.

Examples:

a) VITAL SIGNS: The patient’s heart rate is slightly high, 105bpm.
b) He has a fever with a temperature of 101.6°F.

Assertions: None.

Relations: Is_result_of.

Drug_Ingredient

Definition: All mentions of drug ingredient names included in the text.

Extraction rules: Only generic names should be extracted using

this label. No other posology/drug brand names information should be labeled as Drug.

Examples:

a) Acetaminophen 500mg every day.
b) Nystatin 500mg every 8 hours.
c) He was taking his medication as intended.
d) Tylenol was prescribed for two weeks.

Assertions: None.

Relations: None.

Assertion Labels

Assertion labels are used to indicate an attribute of an entity. The assertion label is placed on top of the entity label, example, Entity Assertion.

The following are considerations when adding assertions:

- Entities should only be assigned one assertion label only.
- For the annotation of Assertion, it should be considered only the information found in the sentence that includes the asserted entity.

Not all the combinations of entity and assertion are possible. A table of all the entities and possible assertions is included at the end of this section.

Present

Definition: Entities referring to the patient that are currently present and not negated.

Extraction rules: Use this assertion label only for entities extracted as Disease_Syndrome_Disorder.

Example:

a) He is a 60 years old gentleman with diabetes Present Disease_Syndrome_Disorder. (Disease_Syndrome_Disorder + Present Assertion).

Absent

Definition: Label added to negated entities.

Extraction rules: Absent entities are found in phrases that include words suchas *no, without, lack,* etc.

Examples:

a) The ultrasound showed that the patient does not have psoriasis Absent Disease_Syndrome_Disorder.(Disease_Syndrome_Disorder Entity + Absent Assertion).
b) She is neither Diabetic Absent Disease_Syndrome_Disorder nor diagnosed with Obesity Absent Disease_Syndrome_Disorder. (Disease_Syndrome_Disorder Entity + Absent Assertion).

Possible

Definition: This label is assigned to entities that are possible but not confirmed.

Extraction rules: Possible entities are found in phrases that include words suchas *maybe, perhaps, could, likely, unlikely, to rule out*, etc.

Example:

a) The patient requires further evaluation to rule out cancer Possible Disease_Syndrome_Disorder.(Disease_Syndrome_Disorder Entity + Possible Assertion).

Family

Definition: This label is assigned to entities linked to a family member of the patient.

Extraction rules: If a medical problem referring to a family member is negated in the text, and both the **Family** assertion and the **Absent** assertion can be applied. Use only the assertion label **Absent,** in that case.

Examples:

a) His father died from colon cancer Family Disease_Syndrome_Disorder. (Disease_Syndrome_Disorder Entity + FamilyAssertion).
b) Her brother died from Stroke Family Disease_Syndrome_Disorder. (Disease_Syndrome_Disorder Entity + FamilyAssertion).

John Snow LABS

Assertion Table

	Present	Absent	Possible	Family
Age	No	No	No	No
Gender	No	No	No	No
Employment	No	No	No	No
Date	No	No	No	No
Disease_Syndrome_Disorder	**Yes**	**Yes**	**Yes**	**Yes**
Modifier	No	No	No	No
VS_Finding	No	No	No	No
Test_Result	No	No	No	No
Drug_Ingredient	No	No	No	No

Relations

Relations are used to link two related entities. To create relations between entities, use the Create Relation button of the annotation tool. Relations are **NOT** created for entities present in different sentences, or are 2 or more sentences apart. Some relations require a relation label. Also, some relations require assignation of direction that is represented by an arrow in the relation section of the annotator tool. A table with all the possible relations is included at the end of this section.

Is_diagnosis_date_of

Definition: This relation is used to associate a Disease_Syndrome_Disorder entity and a Date entity.

Extraction rules: The Disease_Syndrome_Disorder entity and the relevant date associatedwith it are extracted and related using the relation label is_diagnosis_date_of only when the date refers to the diagnosis of the medical problem.

Examples:

a) She was diagnosed with Parkinsons in 1987. **Parkinsons** (Disease_Syndrome_Disorder entity) and **1978** (Date entity) are related with is_diagnosis_date_of label.

Is_modifier_of

Definition: This relation is used to associate a Disease_Syndrome_Disorder entity and a Modifier.

Extraction rules: The Disease_Syndrome_Disorder and the relevant modifier are extractedand related using the relation label is_modifier_of.

Examples:

a) He has been experiencing chronic migraine for five years. **Migraine** (Disease_Syndrome_Disorder entity) and **Chronic** (Modifier entity) are related with is_modifier_of label.

b) Patient with history of recurrent atrial fibrillation. **arrythmias** (Disease_Syndrome_Disorder entity)

and **Recurrent** (Modifier entity) are related with is_modifier_of label.

Is_cause_of

Definition: This relation is used to associate two Disease_Syndrome_Disorder entities when one of them is caused by the other.

Extraction rules: First, extract both Disease_Syndrome_Disorder. Then, create the relation selecting first the cause and then the consequence, and use the label is_cause_of.

Examples:

a) He has esophageal varices secondary to liver cirrhosis. **Esophageal Varices** (Disease_Syndrome_Disorder Entity) and **Liver Cirrhosis** (Disease_Syndrome_Disorder Entity) are related with is_cause_of label, with direction from esophageal varices towards liver cirrhosis.

Is_result_of

Definition: This relation should be used to associate a VS_Finding entity and a Test_Result entity.

Extraction rules: The VS_Finding entity and the relevant Test_Result are extracted and related with the is_result_of label.

Examples:

a) He is febrile with a temperature of 101.6°F. **Temperature** (VS_Finding) and **101.6°F** (Test_Result Entity) are related with is_result_of label.

Relation Table

	Entity 1	Entity 2	Label Needed	Direction Needed
Is_diagnosis_date_of	Disease_Syndrome_Disorder	Date	Yes	No
Is_modifier_of	Disease_Syndrome_Disorder	Modifier	Yes	No
Is_cause_of	Disease_Syndrome_Disorder	Disease_Syndrome_Disorder	Yes	Yes
Is_result_of	VS_Finding	Test_Result	Yes	No

Text Classification

Checkboxes are used to label information at a document level. The information to select an answer should be searched in the whole text. The following checkboxes are included in this project:

Gender

- ❑ Female
- ❑ Male
- ❑ Unknown

Type of Text

- ❑ Surgical Note – Text relating to a surgical procedure
- ❑ Other Clinical Text – Non-surgical text i.e. Medicine, Radiology, Tests/ Test Results, etc.
- ❑ Unknown – Non-medical text or non-classified document due to lack of information

Change Log

Version	Revision Date	Revision Description	Responsible for AG Updates
1.0	June 23rd, 2023	Updates to implement following the consensus meeting – 22.07.2023	@Annotator 1
2.0	June 27th, 2023	Updates to implement following the consensus meeting – 27.07.2023	@Annotator 2