Guidelines for entity annotation in NEREL-BIO dataset

Overview

This documentation describes the guidelines and its rationale for annotating entities in biomedical abstracts of the NEREL-BIO dataset. The NEREL-BIO dataset contains PubMED abstracts devoted to discussion of diseases and disorders and their treatment. Current annotation includes 37 entity types. The entity annotation is the first level of annotation for next relation annotation and entity linking with the UMLS knowledge base. Specific features of NEREL-BIO annotation are as follows:

- Annotation of nested entities, which means that entities can contain each other or intersect – such annotation allows annotating new complex entities discussed in the text and their relations and at the same time allows linking internal known entities to UMLS;
- Entity absent in UMLS can be annotated because they can provide new results of studies, absent concepts, region-specific concept etc. important for understanding the results of scientific studies presented in the abstract;
- Annotated entries can be fragmented (divided into several fragments) but usually we try to avoid the annotation with fragmented entities,
- Entity types of NEREL-BIO include biomedical entities and also general entity types for description of such as ORGANIZATION or COUNTRY important for description of social and organization aspects of disease treatment. General entity types are based on NEREL dataset[1] of news media texts annotated for named entity recognition in the general domain.

General rules for entity annotation:

1) Annotated entity usually should present a single concept from pre-defined types.
2) Annotated entity can be a single token or multitoken sequences. A single token can be a word (noun, adjective or verb) or digit (for numerical entity types). General words with meaning corresponding to entity type sense are not annotated as a single entity (f.e. “disease”, disorder, ‘study”, “organ”)
3) Multiword entity usually should contain 2-4 words of adjectives or nouns without prepositions in Russian or can possibly include “of” preposition in English. Longer

[1] Loukachevitch, N., Artemova, E., Batura, T., Braslavski, P., Denisov, I., Ivanov, V., ... & Tutubalina, E. (2021). NEREL: A Russian Dataset with Nested Named Entities, Relations and Events. In proceedings of RANLP-2021, p. 876–885

sequencies or sequences containing prepositions should be subdivided into shorter spans of entities with the following exceptions:

   a. The abstract contains abbreviation for longer or more complex entity;
   b. The term is a vocabulary entry in referential domain-specific resources (UMLS, medical dictionaries);
   c. FINDING entity can be more complex and longer, has its own rules for annotation.

4) Adjectives within longer terms with the same meaning are nor annotated as depressive in depressive disorder: [depressive disorder]$_{\text{DISO.}}$

List of entities and corresponding UMLS semantic groups and concepts

N	NEREL-BIO entity type	UMLS Semantic GROUP	UMLS concept name and identifier
1	ACTIVITY	B1.1.2 Individual Behavior	Behavior UMLS CUI C0004927
2	ADMINISTRATION_ROUTE	A2.1.4 Functional concept	Drug Administration Routes UMLS CUI C0013153
3	ANATOMY	A1.2 Anatomical structure	Body structure UMLS CUI CC1268086
4	CHEM	A1.4.1 Chemical	Chemical Substance (organic or inorganic) UMLS CUI C0220806
5	DEVICE	A1.3.1 Medical device	Medical devices UMLS CUI C0025080
6	DISO	B2.2.1.2 Pathologic Function	Pathology processes C0677042
7	FINDING	A2.2 Finding	Experimental finding C2825141
8	FOOD	A1.4.3 Food	Food UMLS CUI C0016452
9	GENE	A1.2.3.5 Gene or Genome	Genes UMLS CUI C0017337 and DNA UMLS CUI C0012854

10	INJURY_POISONING	B2.3 3 Injury & Poisoning	Poisoning / injury UMLS CUI C0178314
11	HEALTH_CARE_ACTIVITY	B1.3.1 Health Care Activity	Health Care C0086388
12	LABPROC	B1.3.1.1 Laboratory Procedure	medical screening and diagnostic method UMLS CUI C0679541
13	LIVB	A1.1 Organism	Organism UMLS CUI C0029235
14	MEDPROC	B1.3.1.3 Therapy or Preventive procedure	Therapeutic procedure UMLS CUI C0087111
15	MENTALPROC	B2.2.1.1.1.1 Mental Process	Mental processes UMLS CUI C0025361
16	PHYS	B2.2.1.1 Physiologic Function	Physiological processes UMLS CUI C0031845
17	SCIPROC	B1.3.2 Research Activity	Research Activities UMLS CUI C0242481
General entities from NEREL dataset			
18	AGE	A2.1.3 Quantitative concept	Age UMLS CUI C0001779
19	CITY	A2.1.5.4 Geographical areas	Municipality UMLS CUI C0600182
20	COUNTRY	A2.1.5.4 Geographical areas	Country UMLS CUI C0454664
21	DATE	A2.1.1 Temporal concept	Date in time

			UMLS CUI C0011008
22	DISTRICT	A2.1.5.4 Geographical areas	District UMLS CUI C5447118
23	EVENT	B Event	Environmental Event UMLS CUI C0456156 Traumatic event UMLS CUI C4751223
24	FAMILY	A2.9.3 Family Group	C0015576
25	FACILITY	A2.1.5.4 Geographical areas	Facility (object) UMLS CUI C1547538
26	LOCATION	A2.1.5.4 Geographical areas	Location UMLS CUI C0450429
27	MONEY	A1.3 Manufactured Object	Money UMLS CUI C0870909
28	NATIONALITY	A2.8 Group Attribute	Nationality UMLS CUI C0027473
29	NUMBER	A2.1.3 Quantitative concept	Numbers UMLS CUI C0237753
30	ORDINAL	A2.4 Intellectual product	Ordinal number UMLS CUI C0439080
31	ORGANIZATION	A2.7 Organizations	entity - organization UMLS CUI C1561598
32	PERCENT	A2.1.3 Quantitative concept	Percent (qualifier value) UMLS CUI: C0439165

33	PERSON	A1.1.3.1.1.4.1 Human	Homo sapiens UMLS CUI: C0086418
34	PRODUCT	A1.3 Manufactured Object	Product UMLS CUI: C1514468
35	PROFESSION	A2.9.1 Professional Group	Workers UMLS CUI: C1527116
36	STATE_OR_PROVINCE	A2.1.5.4 Geographical areas	State or province UMLS CUI: C1555317
37	TIME	A2.1.1 Temporal concept	Time UMLS CUI: C0040223

Guidelines for each entity type with examples

ADMINISTRATION_ROUTE

Administration route through which chemical is inserted to the body

Eg : injection, oral intake, IV

- Experimentally [[injected] $_{\textbf{ADMINISTRATION\_ROUTE}}$ into the [anterior chamber]$_{\text{ANATOMY}}$]$_{\textbf{ADMINISTRATION\_ROUTE}}$
- [6]$_{\text{NUMBER}}$ [rabbits]$_{\text{LIVB}}$ received [sterile [intracameral [injections]$_{\textbf{ADMINISTRATION\_ROUTE}}$]$_{\textbf{ADMINISTRATION\_ROUTE}}$]$_{\textbf{ADMINISTRATION\_ROUTE}}$ of [0.2 ml]$_{\text{NUMBER}}$

ACTIVITY

The actions or reactions of a person or an organism

Eg: smoking, contact with animals, to give up smoking, military actions, illegal drug use

ANATOMY

Anatomy, Body parts, organs, structure of organisms and their parts, Body locations and regions,

Eg : Lung, Heart, Dead body, Arms, Legs, right hemisphere, pulmonary, cardiac

- [[lymph]$_{\textbf{ANATOMY}}$ [capillaries]$_{\textbf{ANATOMY}}$]$_{\textbf{ANATOMY}}$,
- [[nerve]$_{\textbf{ANATOMY}}$ endings]$_{\textbf{ANATOMY}}$
- [lower [limb]$_{\textbf{ANATOMY}}$]$_{\textbf{ANATOMY}}$]
- [[coronary]$_{\textbf{ANATOMY}}$ [heart]$_{\textbf{ANATOMY}}$ disease]$_{\text{disso}}$

CHEM

Chemicals and drugs, contains both organic and inorganic, pharmaceuticals medicines

Eg : Uric acid, nitrogen dioxide, sulphur,

- [Asketin [cefuroxime]$_{\textbf{CHEM}}$]$_{\textbf{CHEM}}$ [1 mg/ml]$_{\text{NUMBER}}$, [gentamicin]$_{\textbf{CHEM}}$ [200 µg/ml]$_{\text{NUMBER}}$, [amikacin]$_{\textbf{CHEM}}$ [400 µg/ml]$_{\text{NUMBER}}$, [amphotericin B]$_{\textbf{CHEM}}$ [10 µg/ml]$_{\text{NUMBER}}$
- [35]$_{\text{NUMBER}}$ [patients]$_{\text{PERSON}}$ who received therapy with [Influnet]$_{\textbf{CHEM}}$

DEVICE

Devices. Medical or Research Devices

Eg : Ventilator, prosthesis

- [Super lightweight [polypropylenepolyvinylidenefluoride]$_{CHEM}$ [prosthesis]$_{\mathbf{DEVICE}}$]$_{\mathbf{DEVICE}}$
- [infection of [[vascular]$_{ANATOMY}$ [prostheses]$_{\mathbf{DEVICE}}$]$_{\mathbf{DEVICE}}$]$_{DISO}$

DISO

Particular abnormal condition that negatively affects the structure or function of all or part of an organism, complications, etc.

- [patients]$_{PERSON}$ with [[liver]$_{ANATOMY}$ [metastatic [lesion]$_{\mathbf{DISO}}$]$_{\mathbf{DISO}}$]$_{\mathbf{DISO}}$
- [patients]$_{PERSON}$ of different [age groups]$_{PHYS}$ presenting with [acute [rhinosinusitis]$_{\mathbf{DISO}}$]$_{\mathbf{DISO}}$
- [[Pulmonary]$_{ANATOMY}$ [paecilomycosis]$_{\mathbf{DISO}}$]$_{\mathbf{DISO}}$

We do not annotate general terms such as disease, illness etc as DISO.

Stages of the disease are annotated as DISO, for example:

- [[chronic [[kidney] $_{\mathbf{ANATOMY}}$ disease] $_{\mathbf{DISO}}$] $_{\mathbf{DISO}}$ stage [1] $_{\mathbf{ORDINAL}}$]$_{\mathbf{DISO}}$

FINDING

Indicates result, consequence, effects of some procedures, chemical preparations, disease etc. FINDING entity presents main results of a study in a short form of keywords. It usually consists of two parts: finding markers such as *increase, improvement, efficacy* (or sequence of markers: *efficiency and safety*) and one of NEREL_BIO entities (or their general mentions as disorder) usually to the right from markers. The finding markers can be subdivided in two groups: changes as *increase, improvement, efficacy* and group statistics such as mortality, number of registered cases, etc.

This a new form of keywords announcing main results of the study therefore such FINDING entities are mostly not included in any referential resources.

Eg : laralized increase, improvement of quality of life

- A [significant reduction of [[pain]$_{DISO}$ syndrome]$_{DISO}$]$_{\mathbf{FINDING}}$
- [[high [procoagulant activity]$_{PHYS}$]$_{\mathbf{FINDING}}$ of the [hemostatic]$_{PHYS}$ system]$_{\mathbf{FINDING}}$ was observed during all stages of the study

DO NOT include to FINDING entity extra non-meaningful words such as *allow* or *revealed,* produce etc., without which the meaning of the finding continues to be understandable.

FOOD

Substances taken in by the body to provide nourishment

Eg ***:*** salt, milk, hot meal, fresh water, alcoholic beverages, food fusion for kids

GENE

nucleic acid sequences that function as units of heredity, also comprises chromosomes and DNA.
Eg. KIR gene, PARK2, RASSF1A, allele, CYP2C19

HEALTH_CARE_ACTIVITY

the actions of prevention and management of treatment
Eg.: hospitalization, discharge from hospital, medical evacuation, patient transfer, hospital admission

INJURY_POISIONING

Injury or poisoning corresponds to UMLS semantic type |T037|Injury or Poisoning.
Damage to certain Anatomical parts like fractures, bruises, wounds etc are included in this group as such injury is not a disease.

- [[Lung]$_{\text{ANATOMY}}$ and [left [heart]$_{\text{ANATOMY}}$]$_{\text{ANATOMY}}$ injuries]$_{\textbf{INJURY\_POISIONING*}}$ and [thromboembolism]$_{\text{DISO}}$
- [cracks]$_{\textbf{INJURY\_POISIONING}}$ and complete [[fractures]$_{\textbf{INJURY\_POISIONING}}$ of the [[tracheal]$_{\text{ANATOMY}}$ [cartilages]$_{\text{ANATOMY}}$]$_{\text{ANATOMY}}$]$_{\textbf{INJURY\_POISIONING}}$
-

LABPROC

Laboratory Process corresponds to UMLS semantic type |T059|Laboratory Procedure,
Used to denote the processes that are performed in the lab rather than medical procedures.
Eg : blood biochemical test, immunological test, ECG, ultrasound

- [[ECG]$_{\textbf{LABPROC}}$ monitoring]$_{\textbf{LABPROC}}$ in the [animals]$_{\text{LIVB}}$
- [normalization of the [[lung]$_{\text{ANATOMY}}$ [ultrasound]$_{\textbf{LABPROC}}$]$_{\textbf{LABPROC}}$ pattern]$_{\text{FINDING}}$
- [[blood]$_{\text{ANATOMY}}$ general]$_{\textbf{LABPROC}}$ and [immunological tests]$_{\textbf{LABPROC}}$

LIVB

Living beings, contains all living organism, it is a general class, use this to annotate living beings other than PERSON or PROFESSION
Eg : Rabbits, Virus, Bacteria, Fungi

- [[Fungi]$_{\textbf{LIVB}}$ of the [[Paecilomyces]$_{\textbf{LIVB}}$ genus]$_{\textbf{LIVB}}$]$_{\textbf{LIVB}}$

MEDPROC

Medical Process, contains UMLS semantic types |T061|Therapeutic or Preventive Procedure

Used to denote the medica, surgical, therapeutic procedures

Eg : chemotherapy, cesarean section

- [[aortocoronary]$_{\text{ANATOMY}}$ [bypass [surgery]$_{\textbf{MEDPROC}}$]$_{\textbf{MEDPROC}}$]$_{\textbf{MEDPROC}}$ for [CHD]$_{\text{DISO}}$
- [[tracheal]$_{\text{ANATOMY}}$ [intubation]$_{\textbf{MEDPROC}}$]$_{\textbf{MEDPROC}}$ prior to [death]$_{\text{PHYS}}$
- [Isolated [bronchus]$_{\text{ANATOMY}}$ [resection]$_{\textbf{MEDPROC}}$]$_{\textbf{MEDPROC}}$
- [[mediastinal]$_{\text{ANATOMY}}$ [lymphadenectomy]$_{\textbf{MEDPROC}}$]$_{\textbf{MEDPROC}}$

MENTALPROC

Mental Process, contains UMLS semantic type |T041|Mental Process

Used to denote the mental process or some neurological or mental activity, cognitive process

Eg, Cognitive, emotion, self-esteem

- Predominance of [[volitional]$_{\textbf{MENTALPROC}}$ disorders]$_{\text{DISO}}$
- [Mild [[cognitive]$_{\textbf{MENTALPROC}}$ impairment]$_{\text{DISO}}$]$_{\text{DISO}}$ in [patients]$_{\text{PERSON}}$

PHYS

Physiology of the human body. functions and mechanisms in a living system

Eg : lymphocirculatory, microhemocirculatory, innervation, young

- [impaired [maturation]$_{\textbf{PHYS}}$]$_{\textbf{DISO}}$ in all the study cases of [[vascular]$_{\text{ANATOMY}}$ [aging]$_{\textbf{PHYS}}$]$_{\textbf{PHYS}}$
- [[regeneration]$_{\textbf{PHYS}}$ of [[lung]$_{\text{ANATOMY}}$ [tissue]$_{\text{ANATOMY}}$]$_{\text{ANATOMY}}$]$_{\textbf{PHYS}}$
- [[Liver]$_{\text{ANATOMY}}$ [[enzymes]$_{\text{CHEM}}$ activity]$_{\textbf{PHYS}}$]$_{\textbf{PHYS}}$

Do not annotate indefinite, immeasurable expressions as *level of neurotization* or *state of liver* as PHYS.

SCIPROC

Scientific Process, contains UMLS semantic types |T065|Educational Activity, |T063|Molecular Biology Research Technique, |T062|Research Activity

Used to denote the scientific procedure performed. SCIPROC includes procedures that can't be added to LABPROC or MEDPROC. Also include Questionnaires or some other scales to determine the level of disease.

Eg : Multivariate statistical analysis, multicenter double-blind randomized placebo-controlled study.

- [[Paired t-]$_{\textbf{SCIPROC}}$ and [χ^2-test]$_{\textbf{SCIPROC}}$]$_{\textbf{SCIPROC}}$ was used for [statistical analysis]$_{\textbf{SCIPROC}}$
- [Hamilton [depression]$_{\text{DISEASE}}$ rating scale]$_{\textbf{SCIPROC}}$

- The [[Chronic [Pain]$_{\text{SYMPTOM}}$ Coping Inventory]$_{\textbf{SCIPROC}}$

DO NOT annotate general terms like analysis, study etc. Specific methods or scales should be discussed. DO NOT annotate such expressions as *assessment of cognitive functions as SCIPROC.*

General entities used in NEREL-BIO

Some entities that are annotated in the general domain dataset NEREL-BIO only if they are proper names (COUNTRY, PERSON, ORGANIZATION, FACILITY, etc.), in NEREL-BIO both abstract and concrete mentions are annotated.

AGE

Age of the person, the word 'age' itself is not annotated. Include years in **AGE** (i.e. [90 years]$_{\textbf{AGE}}$ old)

- died on [October 13, 2016]$_{\text{DATE}}$, at the age of [91]$_{\textbf{AGE}}$

CITY

names or subtypes of populated localities (including small towns and villages): Berlin, Moscow, Brazilian cities.

- [Tel Hashomer Hospital]$_{\text{FACILITY}}$ in [Tel Aviv]$_{\textbf{CITY}}$

COUNTRY

Political state, nation, or territory which is controlled including proper names and types of countries: India, European countries, low- and middle-income countries.

DATE

Reference to a date or a period more than a day (except **age**): January, Monday, next year, [seventies] fashion, tomorrow. Extent should include modifiers and prepositions that denote specific time:

Eg : For 2 days, since last week, 1920-x years

DISTRICT

Type of administrative division that is larger or smaller than province, state, or other main administrative subdivisions of a country

- In the town of [Vishnyava]$_{\text{CITY}}$ ([Vishnevo]$_{\text{CITY}}$) of the [Volozhin district]$_{\textbf{DISTRICT}}$ of the [Novogrudok Voivodeship]$_{\text{STATE\_OR\_PROVINCE}}$ of [Poland]$_{\text{COUNTRY}}$

EVENT

EVENT entity is used for labeling such news-related, non-everyday situations as epidemics, military conflicts, tsunamis, etc, mentioned in connection with the spread of diseases or the need for additional medical care. In UMLS the closest concepts to such treatment of events are two concepts: Environmental Event (C0456156) and Traumatic event **(**C4751223).

FAMILY

Used for labeling patient's families.

FACILITY

Names or types of man-made structures: streets, bridges, buildings, monuments. In NEREL_BIO FACILITY entities are mainly abstract, low-cased mentions.

Eg : hospital, camp, round-the-clock hospital

LOCATION

Names or types of geographical locations other than COUNTRY, STATE_OR_PROVINCE, CITY, DISTRICT, also named regions (Middle East), including continents, water objects (seas, oceans,rivers) mountains, etc.

Eg : Europe, Asia, coastal regions, agricultural area, central part of India, Western Pacific

MONEY

Any monetary values including monetary denominations. Only values should be tagged (generic references to money should not marked): [50 yen]

NATIONALITY

Belonging to a particular nation.

- [episiotomies] $_{\text{MEDPROC}}$ among [Vietnamese] $_{\text{NATIONALITY}}$ [women]$_{\text{PERSON}}$ in [Australia]$_{\text{COUNTRY}}$

NUMBER

If a number is used with standard units (km, kg) – the unit is included in the annotation.

A qualifier before the umber is included (more, less).

Prepositions before number if they do not modify the number are not included.

Wordforms person, copies, pieces (persons, items) are quite frequent with numbers, but not included.

Numbers are not annotated within time, date, money labels because they are also numerical in nature.

- [juvenile [myoclonic [epilepsy]$_{DISO}$]$_{DISO}$]$_{DISO}$ ([JME]$_{DISO}$) - in [9]$_{\mathbf{NUMBER}}$ of [47]$_{\mathbf{NUMBER}}$ ([19.1%]$_{PERCENT}$)

Numerical construction as 1-2 are annotated as a single unit [1-2] $_{\mathbf{NUMBER}}$

ORDINAL

A number designating a place in an ordered sequence.

Eg : first group, [21st]$_{\mathbf{ORDINAL}}$ century

Ordinal constructions as I-II are annotated as a single entity.

ORGANIZATION

Names or types of organizations (companies, governmental organizations, educational institutions, sport teams).

Eg : medical institution, World Health Organization, Ministry of Health, higher education institution, psychotherapeutic service

PERCENT

Any percentage with percent symbol or word "percent". Preposition before percent is not annotated. A qualifier before the percent mention is included (more, less).

- [LTG]$_{CHEM}$, [TPM]$_{CHEM}$ and [CBZ]$_{CHEM}$ in [patients]$_{PERSON}$ with other [epileptic syndromes]$_{SYMPTOMS}$ (non [IGE]$_{DISEASE}$): [4.9%]$_{\mathbf{PERCENT}}$, [4%]$_{\mathbf{PERCENT}}$ and [3.7%]$_{\mathbf{PERCENT}}$, respectively

PERSON

In contrast, to the general domain in NEREL-BIO PERSON is an abstract noun, written lowercase, such as patients, control group, children. Names of persons within disease mentions are not annotated.

PRODUCT

Any product having a model name and possibly model number.

- results of [spiral [computed [tomography]$_{LABPROC}$]$_{LABPROC}$]$_{LABPROC}$ with the use of the [Solid Works 2012]$_{\mathbf{PRODUCT}}$ software package

PROFESSION

includes occupations, posts, professional titles of specific persons or occupational groups. Such words as *author*, *creator*, *developer* are not annotated as PROFESSION because it is more appropriate to describe this information via relations - this is so-called relative nouns.

Eg : physician, expert, medical personnel, surgeon, physician, US president, opera soloists

STATE_OR_PROVINCE

Main administrative subdivision of a country

Eg :Moscow region, Bashkir republic

TIME

Specific moment of the day (in English with a.m. or p.m., which included in the annotation). Other times of the day (morning) are also annotated. Short interval less than 24 hours are also annotated as TIME

- [[Pain]$_{diso}$ threshold]$_{PHYS}$ was measured before and [15]$_{\mathbf{TIME}}$, [30]$_{\mathbf{TIME}}$, [45]$_{\mathbf{TIME}}$ and [60 minutes]$_{\mathbf{TIME}}$ after [[oral]$_{ANATOMY}$ intake]$_{ADMINISTRATION\_ROUTE}$