Annotation Guideline

Motivation

The development of named entity recognition (NER) and event extraction (EE) research heavily depends on the availability of human-annotated corpus. In the biomedical field, NER and EE-related corpus have been developed primarily from experimental literature. The experimental phase can be generally divided into three stages: cells, animal experiments, and clinical trials. Compared to cell experiments, animal research requires further due diligence since they should consider ethical guidelines and have extensive resource requirements. More importantly, before moving on to clinical trials with human participants, animal research serves as a significant step to evaluate safety and efficacy. Therefore, thorough investigations of previous sutdies are essential to design animal experiments, verifying information such as species, dosage, and durations, and the relations between them.

However, most previous datasets contains literature that is either limited to the cell experimental stage (Kim et al., 2011; Pyysalo et al., 2013; Ohta et al., 2013) or does not specify a concrete experimental stage (Pyysalo et la., 2012). An entity and event scheme that do not align with the dataset scope make it difficult to identify specific event triggers and their associated arguments, which are prevalent in animal experiments. Therefore, we introduce **AniEE**, a NER and EE dataset focused on animal experiments.

Data Sampling

As described in *Section 3.3.1 Annotation Procedure* of the paper, prior to performing the actual annotation process, a pilot annotation was conducted to train the annotators due to their varying levels of domain knowledge. It was performed to apply our newly defined scheme on 10 pieces of literature (PMID: 12201487, 15975645, 10473104, 14597675, 12154044, 19139005, 18841154, 11027431, 10589785, 12766067), which are extracted if they are related to the animal experiment stage from the MLEE corpus (Pyysalo et la., 2012), a publicly available biomedical event extraction dataset.

During the pilot annotation, we refined the annotation guideline that we had drafted by referencing previous studies (Li et al., 2016; Luo et al., 2022). To improve the guideline, we compared and discussed the annotation results in collaboration with domain experts (annotators).

Then, we annotated the raw corpus with the updated guideline, which was collected from PubMed, a widely-used online database containing a vast collection of biomedical literature. To crawl a diverse set of animal experimental literature as the raw corpus, we defined the search terms with domain experts. Medical subject headings (MesH), which serves as a

hierarchical search index for the PubMed database, were used to determine the search term. First, we included [MesHTerms]AnimalDiseaseModel OR [MesHTerms]Animals in the search term to retrieve literature that performed animal experiments. Then, to obtain a literature collection on a wide range of topics, we set the search terms of [MesHSubheading]Physiopathology OR [MesHSubheading]Chemistry to include the literature wich falls under physiopahtology or chemistry as broad categories, while we excluded review articles. Thus, the final search terms consisted of ([MesHTerms]AnimalDiseaseModel OR [MesHTerms]Animals) AND ([MesHSubheading]Physiopathology OR [MesHSubheading]Chemistry) AND (Not Review). We collected the titles and abstracts of the literature from the top search results.

We then excluded some of them without the direct involvement of animal experiments, resulting in a total of 350 literature.

Literature exclusion criteria:

If any of 1), 2), or 3) is applicable, the literature is excluded.

1) If none of the mentions corresponding to SampleName or SampleAdministration mention exist
2) If none of the mentions corresponding to AnimalSubject or AnimalStrain exist
3) If none of the mentions corresponding to Molecular Biomarker exist

General Rule

- Annotate all the spans of all the 15 concept types (12 entities & 3 events).
- The full text can be accessed to clarify the concept spans.
- **The abbreviation and its long form should be annotated separately**. For instance, annotate both “Sprague-Dawley Rats” and “SDR” in the text “Sprague-Dawley Rats (SDR)”.
- **Annotate minimum necessary text spans to define the concept types.**
- **Annotate all mentions of the concept types, regardless of whether they are highlighted in the literature.**

Entity Type Rule

SampleName

- Sample refers to a material(compound or substance) addressed in animal experiment.
- Annotate both main effector, inducer, and inhibitor.
    - Main effector : A compound or substance being mainly addressed in the literature
    - Inducer : An agent that initiates or enhances the expression of a specific gene, pathway, or particular activity
    - Inhibitor : An agent that suppresses or reduces the expression of a specific gene, pathway, or particular activity
- Annotate the most specific Sample mention. For instance, annotate “Benzamide derivatives”, instead of “Benzamide”.
- Annotate chemical form.

SampleType

- SampleType refers to a form or preparation of the sample (SampleName) (e.g. “extract”, “oil”, “powder”, “decoction”).

Dosage

- Dosage refers to an amount of sample (SampleName) administered to animals.
- Annotate the mention written as either a number or a character to indicate Dosage.
- Annotate each one if multiple dosages are present. For instance, in the text “3, 5 mg”, annotate “3” and “5” respectively.

DosageUnit

- DosageUnit refers to the unit noun corresponding to Dosage (e.g. “mg/kg”, “mg kg”, “mg/ml”, “mg/L”, “μg/kg”, “μM”, “mg/Kg”, “mg/ml”).
- DosageUnit refers to a unit for a total amount if DosageDayUnit does not exist.

DosageDayUnit

- DosageDayUnit refers to the time unit noun that occurs with DosageUnit (e.g. “day-1”, “/d”, “/day”).
- Annotate the most specific DosageDayUnit mention including special characters such as “/”.

Duration

- Duration refers to a total period of sample administration to animals.
- Duration doesn’t refer to the period of animal handling or disease induction.
- Annotate the mention written as either a number or a character to indicate Duration.

DurationUnit

- DurationUnit refers to the unit noun corresponding to Duration (e.g. “day”, “days”, “week”, “weeks”).

DosageFrequency

- DosageFrequency refers to the actual count that the sample (SampleName) is administered. For instance, in the text “3 every 2 weeks”, annotate “3”.
- Annotate the mention written as either a number or a character to indicate DosageFrequency.

DosageFrequencyCycle

- DosageFrequencyCycle refers to the administration cycle associated with DosageFrequency. For instance, in the text “3 every 2 weeks”, annotate “every 2 weeks”.
- Annotate the most specific mention, including a number and a unit noun.

AnimalSubject

- AnimalSubject refers to the common name of species being experimented on (e.g. “rat”, “rats”, “mouse”, “mice”, “murine”).

AnimalStrain

- AnimalStrain refers to the specific scientific name of species being experimented on (e.g. “Sprague-Dawley”, “Wistar albino”, “Wister”, “CS57BL”, “BALB/c”).
- Annotate the most specific AnimalStrain mention. For instance, for "C57BL/6J" and "C57BL/6N", annotate “C57BL/6J” and "C57BL/6N" respectively, not “C57BL”.

AnimalSex

- AnimalSex refers to the sex of AnimalSubject (or AnimalStrain) (e.g. “female”, “male”).

Anatomy

- Annotate all mentions related to the body parts, including organ system, organ, tissue, body fluid, cell, cellular component, organ function, pathological formation.
- Annotate the most specific mention found in a dictionary of expressions annotated with body-related entity types from the MLEE corpus (Pyysalo et la., 2012).

- Annotate five senses (“vision”, “auditory”, “olfactory”, “gustatory”, and “haptic”).

Molecular Biomarker

- Molecular Biomarker refers to a quantitative or qualitative measurement indicator of cellular-level biological process.
- Annotate the most specific Molecular Biomarker mention found in MarkerDB, an online database of molecular biomarkers (https://markerdb.ca).
- Annotate chemical form.

Response

- Response refers to a physiological change or response associated with Molecular Biomarker.
- Annotate minimum necessary text spans to define Response, considering morphological variations. Prioritize noun. For instance, in the text “antitumor effects”, “antitumor” is only annotated, but in the text “additive effect”, “additive effect” is annotated.
- DO NOT annotate solely general terms (e.g. “effect”).

DiseaseName

- Annotate the most specific Disease mention found in a combination of MesH (https://meshb.nlm.nih.gov/search) and DiseaseOntology (https://disease-ontology.org/).
- Annotate a symptom to disease.
- DO NOT annotate general terms (e.g. “stress”, “pain”).
- Annotate the complete disease mention if it’s included in the DiseaseOntology, even if it contains general terms. For instance, DO NOT annotate “stress” but annotate “acute stress disorder”.

- To emphasize, annotate all disease mentions beyond the main disease(s) addressed in the literature. Annotate all occurrences of the same disease mention including abbreviations.
- Annotate minimum necessary text spans for a disease. For instance, annotate “hypertension” instead of “sustained hypertension”.

Relation Annotation

Annotate the relation between numeric entity types and temporary unit noun entity types.

- Dosage - DosageUnit
- Dosage - DosageUnit & DosageDayUnit
- DosageFrequency - DosageFrequencyCycle
- Duration - DurationUnit

Event Trigger Rule

SampleAdministration

- SampleAdministration refers to administration of a specific sample to the experimental subject, including injection, oral administratikon, and topical application.
- Annotate abbreviations. For instance, annotate “i.v.”, “i.p.”.
- Annotate the adverb mention (e.g. “orally”).
- Annotate general terms (e.g. “injection”) if body part is explicitly described as to where it is administered.
- Annotate general terms if the most-specific SampleAdministration mention (e.g. “i.v.”) doesn’t exist.
- Annotate the composite verb mention that combines a body part and an act of administration.
- Annotate “topical”.

Positive Regulation

- PositiveRegulation refers to stimulation of a biological process or system in animals that increases the activity, expression, or response of a particular target.
- Annotate the verb mention based on the positiveness of it itself, not the positiveness of the literature result. For instance, in the text “Cancer cells have proliferated.”, annotate it as PositiveRegulation since the mention “prolifereated” itself implies stimulation even though the entire text means a negative outcome.
- Annotate the verb mention with contextual meaning of stimulation if it is a neutral word (e.g. “resulted”).
- Annotate the verb mention with preposition if it is written in passive form.

Negative Regulation

- NegativeRegulation refers to suppression or inhibition of a biological process or system in animals, resulting in reduced activity, expression, or response of a specific target.
- Annotate the verb mention based on the negativeness of it itself, not the negativeness of the literature result. For instance, in the text “Cancer cells have decreased.”, annotate it as NegativeRegulation since the mention “decreased” itself implies supression even though the entire text means a positive outcome.
- Annotate the verb mention with contextual meaning of suppression or inhibition if it is a neutral word (e.g. “resulted”).
- Annotate the verb mention with preposition if it is written in passive form.

Event Argument Rule

- Annotate the relation label of the mention across multiple sentences instead of just one sentence.
- Annotate the relation label of the mention closest to the event trigger mention if multiple mentions for an argument exist.

SampleAdministration

- Object : A material (*SampleName*) which is used for an event SampleAdministration.
- Subject : An animal experimental subject (*AnimalSubject or AnimalStrain*) of an event SampleAdministration.
- Site : A body region (*Anatomy*) where an event SampleAdministration occurs.
- Amount : Quantity measurement (*Dosage*) of a sample (*SampleName, Object*).
    - Annotate multiple Amount argument relation if there are multiple Dosages.
- Schedule : A time frame (*Duration, DosageFrequency*) of an event SampleAdministration.

PositiveRegulation/NegativeRegulation

- Cause : Attribute that influences the modifications of the target factor.
    - An entity type candidate : *SampleName, MolecularBiomarker, Response, DiseaseName, PositiveRegulation, NegativeRegulation*

- Object : Physiological parameters affected by an event SampleAdministration
    - An entity type candidate : *SampleName, MolecularBiomarker, Response, DiseaseName, PositiveRegulation, NegativeRegulation*

- Site : Physiological region (*Anatomy*) where an Object argument is observed.

Reference

[1] Kim, J. D., Wang, Y., Takagi, T., & Yonezawa, A. (2011, June). Overview of genia event task in bionlp shared task 2011. In *Proceedings of BioNLP shared task 2011 workshop* (pp. 7-15).

[2] Pyysalo, S., Ohta, T., & Ananiadou, S. (2013, August). Overview of the cancer genetics (cg) task of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop* (pp. 58-66).

[3] Pyysalo, S., Ohta, T., Miwa, M., Cho, H. C., Tsujii, J. I., & Ananiadou, S. (2012). Event extraction across multiple levels of biological organization. *Bioinformatics*, *28*(18), i575-i581.

[4] Ohta, T., Pyysalo, S., Rak, R., Rowley, A., Chun, H. W., Jung, S. J., ... & Tsujii, J. I. (2013, August). Overview of the pathway curation (PC) task of bioNLP shared task 2013. In *Proceedings of the BioNLP shared task 2013 workshop* (pp. 67-75).

[5] Luo, L., Lai, P. T., Wei, C. H., Arighi, C. N., & Lu, Z. (2022). BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, *23*(5), bbac282.

[6] Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C. H., Leaman, R., ... & Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, *2016*.