Appendix

Chaitanya Kulkarni[1], **Wei Xu**[1], **Alan Ritter**[1], and **Raghu Machiraju**[1]

[1]Department of Computer Science and Engineering, Ohio State University
{kulkarni.132,xu.1265,ritter.1492,machiraju.1}@osu.edu

1 Annotation Rules

The wet lab protocol dataset annotations rules were designed primarily to provide a simple description of the various actions and their arguments in protocols so that it could be more accessible and be effectively used by non-biologists who may want to use this dataset for various natural language processing tasks such as action trigger detection or relation extraction. In the following sections we describe the guidelines that were used in annotating the 622 protocols as we explore the actions, entities and relations that were chosen to be labelled in this dataset.

2 WLP Dataset Curation

The dataset was curated using openly accessible repositories of protocols on platforms such as https://www.protocols.io and http://www.openwetware.org/. The protocols cover a large spectrum of experimental biology, including neurology, epigenetics, metabolomics, cancer/stem cell biology, etc., as seen in Table 2.

3 Actions

Under a broad categorization, Action is a process of doing something, typically to achieve an aim. In the context of wet lab protocols, action words in a sentence or a step are deliberate but short descriptions of a task tying together various other entities in a meaningful way. A fair assumption about action is that they are all verbs. If we were to quickly run a POS tagger, a large percentage of the total action words in our corpus would be tagged as verbs, However some could be mistaken for nouns like "Centrifuge". Some examples of action words, (categorized using GENIA POS tagger), are present in Table 1 along with their frequencies, to quickly demonstrate the difficulty in extracting actions from sentences.

POS tag (freq.)	Top 3 examples
VB (9345)	Add(1404), Incubate(638), Remove(396)
VBG (755)	adding(112), inverting(89), pipetting(34)
VBN (727)	added(43), stored(38), incubated(38)
VBP (512)	Do(80), mix(38), pour(33)
VBD (147)	resuspend(25), put(20), kept(8)
VBZ (44)	remains(5), covers(4), washes(3)
NN (4248)	Centrifuge(324), Transfer(301), Place(215)
NNP (1551)	Mix(335), Wash(277), Vortex(114)
NNS (80)	washes(9), to(7), dilutions(4)
JJ (576)	dry(66), Apply(26), decant(23)
OTHER (1080)	not(111), off(110), up(105)

Table 1: Frequency of different part-of-speech (POS) tags for action words. Majority of the action words either fall under the verb POS tags (VBs 60.48%) or nouns (NNs 30.84%). The GENIA POS tagger is under-identifying verbs in the wet lab protocols, tagging some as adjectives (JJ).

4 Entity

After an extensive annotation process (described in the main paper), we broadly classify various entities commonly seen in protocols under 17 tags. Each of the entity tags were designed to keep a short span length, with the average number of words per entity tag being 1.6. For example, `Concentration` tags are often very short: *60%* *10x*, *10M*, *1 g/ml*, while the `Method` tag has the longest average span of 2.232 words with examples such as *rolling back and forth between two hands* (as seen in Figure 1). The methods in wet lab protocols tend to be descriptive, which pose distinct challenges from existing named entity extraction research in the medical and other domains.

4.1 Object Based Entities

Reagent: A substance or mixture for use in any kind of reaction in preparing a product because of its chemical or biological activity. There are varying definitions of the term reagent. We choose to define it as a label that encompasses all types of chemical, biological and physical substance.

Location: Containers for reagents or other physical entities. They lack any operation capabili-

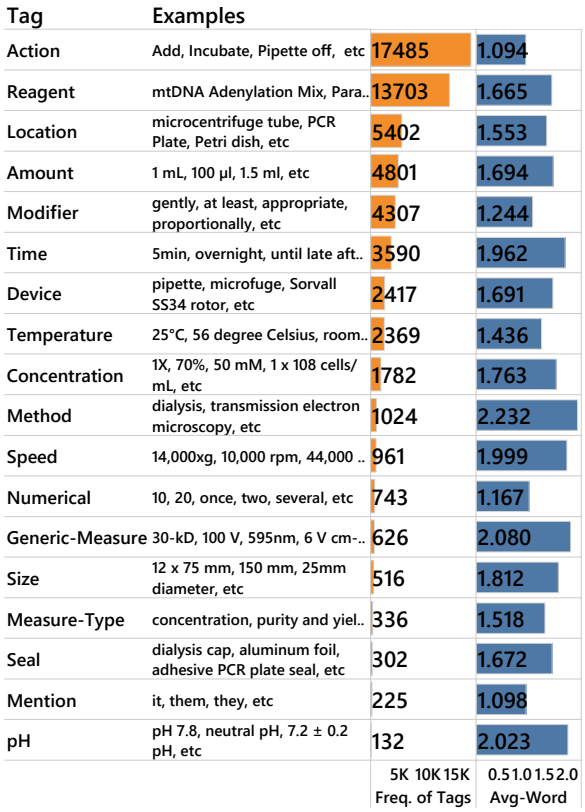


Figure 1: Examples, Frequency and Avg-Word for actions and entities.

ties other than acting as a container. These could be laboratory glassware or plastic tubing meant to hold chemicals or biological substances.

Device: A machine capable of acting as a container as well as performing a specific task on the objects that it holds. A device and a location are similar in all aspects except that a device performs a specific set of operations on its contents, usually illustrated in the sentence itself, or sometimes implied.

Seal: Any kind of lid or enclosure for the location or device. It could be a cap, or a membrane that actively participates in the protocol action, and hence is essential to capture this type of entity.

4.2 Measure Based Entities

Amount: The amount of any Entity being used in any given step, in terms of weight or volume.

Concentration: Measure of the relative proportions of two or more quantities in a mixture. Usually in terms of their percentages by weight or volume.

Time: In the context of wet lab protocols, its to measure the duration for a specific action described in a single step or steps, typically in secs, min, days, or weeks.

Temperature: Any temperature mentioned in degree Celsius, Fahrenheit, or Kelvin.

Method: A word or phrase used to concisely define the procedure to be performed in association with the chosen action verb. Its usually a noun, could also be a passive verb.

Speed: any measure that represents rotation per min for centrifuges.

Numerical: A generic tag for a numerical that doesn't fit time, temp, etc and which isn't accompanied by its unit of measure. Just a number.

Generic-Measure: Any measures that doesn't fit the list of defined measures in this entity list.

Size A measure of the dimension of the object. Like length, or area or thickness of the entity.

Measure-Type: A generic tag to mark the type of numerical it's linked to. Occationally, certain steps in protocol will require you to actually perform a specific type of measurement. This tag is used to highlight which measurement is required to be made by the given action.

pH: measure of acidity or alkalinity of a solution.

4.3 Parts of Speech based Entities

Modifier: It's a word or a phrase that acts as an additional description of the entity it's trying to modify. This additional description must be critical to accurately capture the task instructed in the step. For example *quickly mix* vs *slowly mix* are clearly two different actions, informed by their modifiers "quickly" or "slowly" respectively.

Mention: Marks words that refer to the same object mentioned earlier in the sentence. The attempt is to have a combined tag to mark words that refer to entities first mentioned earlier (anaphor), mentioned later(cataphor) and referring two or more entities (split antecedents).

5 Relations

5.1 Action Relations (Action - Entity)

Acts-On: Links the reagent, or location that the action "acts on", typically linking the direct objects in the sentence to the action.

Creates: This relation marks the physical entity that the action creates.

Site: A Link that associates a Location or Device to an action. It indicate that the Device or Location is the site where the action verb performs. It is also used as a way to indicate which entity will finally hold/contain the creation of an action.

Protocol Category	Count	Avg no. Sent.	Avg. no. Words	Avg. no. Entities	Avg. no. Relations	Avg. no. Actions
molecular biology	186	27.42	338.1	85.25	84.20	35.77
microbiology	105	22.07	328.9	74.46	71.71	27.89
cell biology	94	19.23	236.7	61.09	60.95	23.93
Plant biology	48	17.17	219.96	44.67	43.85	20.44
Immunology	79	25.92	339.5	83.17	78.24	32.68
chemical biology	110	14.37	188.3	46.4	47.45	19.01

Table 2: Statistics of our wet lab protocol corpus by protocol category.

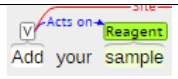
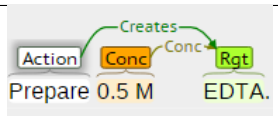
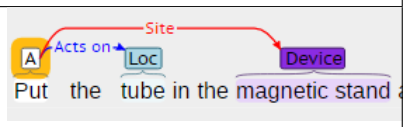
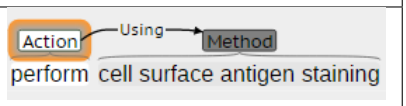
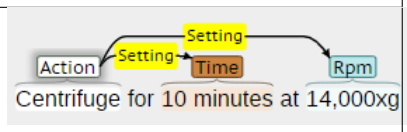
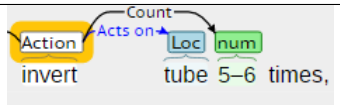
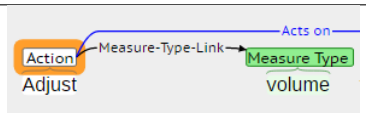
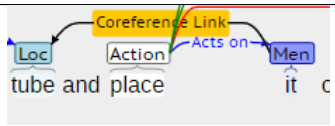
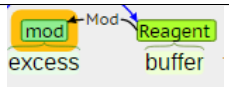
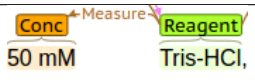
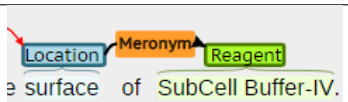
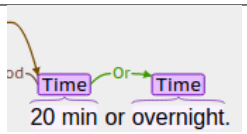
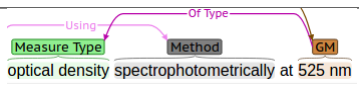
Label	Syntax/Rules	Example
Acts-on	Action ⇒ Reagent | Location | Mention | Device | Seal	
Creates	Action ⇒ Reagent | Mention	
Site	Action ⇒ Location | Device | Mention | Reagent	
Using	Action ⇒ Method | Action | Seal | Device | Mention | Reagent | Location	
Setting	Action | Device | Modifier ⇒ Method | Action | Seal | Device | Mention | Reagent | Location	
Count	Action ⇒ Numerical	
Measure-Type-Link	Action ⇒ Measure-Type	
Coreference	Mention ⇒ [Every other entity]	
Mod-Link	[Every Entity or Action] ⇒ Modifier	
Measure	Reagent | Location | Device | Mention | Seal ⇒ Amount | Numerical | Size | Concentration | Generic-Measure | pH	
Meronym	Reagent | Location | Device | Mention | Seal ⇒ Reagent | Location | Device | Mention | Seal	
Or	[All Entities or Action] ⇒ [All Entities or Action]	
Of-Type	Generic-Measure | Numerical ⇒ Measure-Type	

Table 3: Relations along with their rules and examples

Using: Any entity that the action verb makes use of is linked with this relation. Any entity that the action verb utilizes to perform the action.
Setting: Any measure type entity that is being used to set a device is linked to the action that is attempting to use that numerical.
Count: A Numerical entity that represents the number of times the action must take place is linked using this link.
Measure Type Link: Associates the action to the Measure Type entity that the Action is instructing to measure.

5.2 Binary Relations (Entity - Entity)

Coreference: A link that associates two phrases when those two phrases in a text refer to the same entity.
Mod Link: A Modifier entity is linked to any entity that it is attempting to modify using this relation.
Settings: Links devices to their settings directly, only if there is no action word associated with making those settings.
Measure: A link that associates the various numerical measures to the entity its trying to measure directly.
Meronym: linking reagents, location or device with a prerequisite of materials already contained in the said reagent, location or device.
Or: Allows you to chain multiple entities where either of them can be used for a given link.
Of-Type: used to specify the Measure-Type of a Generic-Measure or a Numerical, if the sentence contains this information.