An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols

Chaitanya Kulkarni, Wei Xu, Alan Ritter, Raghu Machiraju
Department of Computer Science and Engineering
Ohio State University
{kulkarni.132,xu.1265,ritter.1492,machiraju.1}@osu.edu

Abstract

We describe an effort to annotate a corpus of natural language instructions consisting of 622 wet lab protocols to facilitate automatic or semi-automatic conversion of protocols into a machine-readable format and benefit biological research. Experimental results demonstrate the utility of our corpus for developing machine learning approaches to shallow semantic parsing of instructional texts. We make our annotated Wet Lab Protocol Corpus available to the research community.[1]

Isolation of temperate phages by plaque agar overlay
1. Melt soft agar overlay tubes in boiling water and place in the 47C water bath.
2. Remove one tube of soft agar from the water bath.
3. Add 1.0 mL host culture and either 1.0 or 0.1 mL viral concentrate.
4. Mix the contents of the tube well by rolling back and forth between two hands, and immediately empty the tube contents onto an agar plate.
5. Sit RT for 5 min.
6. Gently spread the top agar over the agar surface by sliding the plate on the bench surface using a circular motion.
7. Harden the top agar by not disturbing the plates for 30 min.
8. Incubate the plates (top agar side down) overnight to 48 h.
9. Temperate phage plaques will appear as turbid or cloudy plaques, whereas purely lytic phage will appear as sharply defined, clear plaques.

Figure 1: An example wet lab protocol. The first seven steps are imperative sentences, and the last sentence describes the end results and their subsequent utilization.

1 Introduction

As the complexity of biological experiments increases, there is a growing need to automate wet laboratory procedures to avoid mistakes due to human error and also to enhance the reproducibility of experimental biological research (King et al., 2009). Several efforts are currently underway to define machine-readable formats for writing wet lab protocols (Ananthanarayanan and Thies, 2010; Soldatova et al., 2014; Vasilev et al., 2011). The vast majority of today's protocols, however, are written in natural language with jargon and colloquial language constructs that emerge as a byproduct of ad-hoc protocol documentation. This motivates the need for machine reading systems that can interpret the meaning of these natural language instructions, to enhance reproducibility via semantic protocols (e.g. the Aquarium project) and enable robotic automation (Bates et al., 2016) by mapping natural language instructions to executable actions.

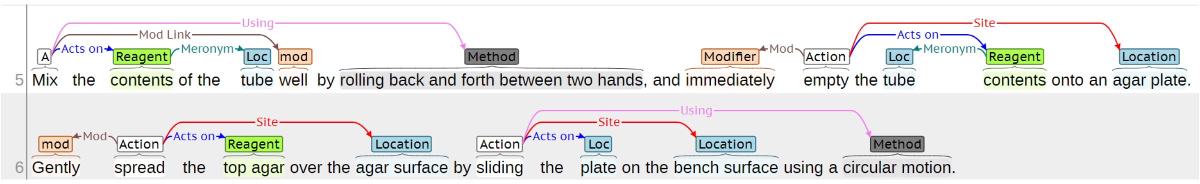
In this study we take a first step towards this goal by annotating a database of wet lab protocols with semantic actions and their arguments; and conducting initial experiments to demonstrate its utility for machine learning approaches to shallow semantic parsing of natural language instructions. To the best of our knowledge, this is the first annotated corpus of natural language instructions in the biomedical domain that is large enough to enable machine learning approaches.

There have been many recent data collection and annotation efforts that have initiated natural language processing research in new directions, for example political framing (Card et al., 2015), question answering (Rajpurkar et al., 2016) and cooking recipes (Jermsurawong and Habash, 2015). Although mapping natural language instructions to machine readable representations is an important direction with many practical applications, we believe current research in this area is hampered by the lack of available annotated corpora. Our annotated corpus of wet lab protocols could enable further research on interpreting natural language instructions, with practical applications in biology and life sciences.

Prior work has explored the problem of learning to map natural language instructions to actions, often learning through indirect supervision

[1] The dataset is available on the authors' websites.

Proceedings of NAACL-HLT 2018, pages 97–106
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics



5 Mix the contents of the tube well by rolling back and forth between two hands, and immediately empty the tube contents onto an agar plate.

6 Gently spread the top agar over the agar surface by sliding the plate on the bench surface using a circular motion.

Figure 2: Example sentences (#5 and #6) from the lab protocol in Figure 1 as shown in the BRAT annotation interface.

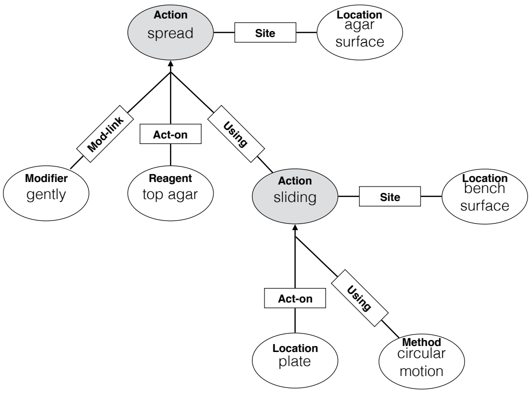


Figure 3: An action graph can be directly derived from annotations as seen in Figure 2 (example sentence #6) .

to address the lack of labeled data in instructional domains. This is done, for example, by interacting with the environment (Branavan et al., 2009, 2010) or observing weakly aligned sequences of instructions and corresponding actions (Chen and Mooney, 2011; Artzi and Zettlemoyer, 2013). In contrast, we present the first steps towards a pragmatic approach based on linguistic annotation (Figure 3). We describe our effort to exhaustively annotate wet lab protocols with actions corresponding to lab procedures and their attributes including materials, instruments and devices used to perform specific actions. As we demonstrate in §6, our corpus can be used to train machine learning models which are capable of automatically annotating lab-protocols with action predicates and their arguments (Gildea and Jurafsky, 2002; Das et al., 2014); this could provide a useful linguistic representation for robotic automation (Bollini et al., 2013) and other downstream applications.

2 Wet Lab Protocols

Wet laboratories are laboratories for conducting biology and chemistry experiments which involve chemicals, drugs, or other materials in liquid solutions or volatile phases. Figure 1 shows one representative wet lab protocol. Research groups around the world curate their own repositories of protocols, each adapted from a canonical source and typically published in the Materials and Method section at the end of a scientific article in biology and chemistry fields. Only recently has there been an effort to gather collections of these protocols and make them easily available. Leveraging an openly accessible repository of protocols curated on the *https://www.protocols.io* platform, we annotated hundreds of academic and commercial protocols maintained by many of the leading bio-science laboratory groups, including Verve Net, Innovative Genomics Institute and New England Biolabs. The protocols cover a large spectrum of experimental biology, including neurology, epigenetics, metabolomics, cancer and stem cell biology, etc (Table 1). Wet lab protocols consist of a sequence of steps, mostly composed of imperative statements meant to describe an action. They also can contain declarative sentences describing the results of a previous action, in addition to general guidelines or warnings about the materials being used.

3 Annotation Scheme

In developing our annotation guidelines we had three primary goals: (1) We aim to produce a semantic representation that is well motivated from a biomedical and linguistic perspective; (2) The guidelines should be easily understood by annotators with or without biology background, as evaluated in Table 3; (3) The resulting corpus should be useful for training machine learning models to automatically extract experimental actions for downstream applications, as evaluated in §6.

We utilized the EXACT2 framework (Soldatova et al., 2014) as a basis for our annotation scheme. We borrowed and renamed 9 object-based entities from EXACT2, in addition, we created 5 measure-based (NUMERICAL, GENERIC-MEASURE, SIZE, PH, MEASURE-TYPE) and 3 other (MENTION, MODIFIER, SEAL) entity types. EXACT2 connects the entities directly to the action without

Protocol Category	Count	avg #Sentences	avg #Words	avg #Entities	avg #Relations	avg #Actions
molecular biology	186	27.42	338.06	85.25	84.20	35.77
microbiology	105	22.07	328.94	74.46	71.71	27.89
cell biology	94	19.23	236.74	61.09	60.95	23.93
Plant biology	48	17.17	219.96	44.67	43.85	20.44
Immunology	79	25.92	339.58	83.17	78.24	32.68
chemical biology	110	14.37	188.30	46.40	47.45	19.01

Table 1: Statistics of our Wet Lab Protocol Corpus by protocol category.

	Total	per Protocol	per Sentence
# of sentences	13679	21.99	–
# of words	177770	285.80	12.99
# of entities	43236	69.51	3.16
# of relations	42425	68.21	3.10
# of actions	17485	28.11	1.28

Table 2: Statistics of the Wet Lab Protocol Corpus.

Annotators	Entities+Actions	Relations
Biologist-Linguist	0.7600	0.6084
Biologist-Other	0.7621	0.6619
Linguist-Other	0.7574	0.6753
all 4 coders	0.7599	0.6625

Table 3: Inter-annotator agreement (Krippendorff's α) between annotators with biology, linguistics and other backgrounds.

describing the type of relations, whereas we defined and annotated 12 types of relations between actions and entities, or pairs of entities (see Appendix for a full description).

For each protocol, the annotators were requested to identify and mark every span of text that corresponds to one of 17 types of entities or an action (see examples in Figure 2). Intersection or overlap of text spans, and the subdivision of words between two spans were not allowed. The annotation guideline was designed to keep the span short for entities, with the average length being 1.6 words. For example, CONCENTRATION tags are often very short: *60% 10x*, *10M*, *1 g/ml*. The METHOD tag has the longest average span of 2.232 words with examples such as *rolling back and forth between two hands*. The methods in wet lab protocols tend to be descriptive, which pose distinct challenges from existing named entity extraction research in the medical (Kim et al., 2003) and other domains. After all entities were labelled, the annotators connected pairs of spans within each sentence by using one of 12 directed links to capture various relationships between spans tagged in the protocol text. While most protocols are written in scientific language, we also observe some non-standard usage, for example using *RT* to refer to *room temperature*, which is tagged as TEMPERATURE.

4 Annotation Process

Our final corpus consists of 622 protocols annotated by a team of 10 annotators. Corpus statistics are provided in Table 1 and 2. In the first phase of annotation, we worked with a subset of 4 annotators including one linguist and one biologist to develop the annotation guideline for 6 iterations. For each iteration, we asked all 4 annotators to annotate the same 10 protocols and measured their inter-annotator agreement, which in turn helped in determining the validity of the refined guidelines. The average time to annotate a single protocol of 40 sentences was approximately 33 minutes, across all annotators.

4.1 Inter-Annotator Agreement

We used Krippendorff's α for nominal data (Krippendorff, 2004) to measure the inter-rater agreement for entities, actions and relations. For entities, we measured agreement at the word-level by tagging each word in a span with the span's label. To evaluate inter-rater agreement for relations between annotated spans, we consider every pair of spans within a step and then test for matches between annotators (partial entity matches are allowed). We then compute Krippendorff's α over relations between matching pairs of spans. Inter-rater agreement for entities, actions and relations is presented in Figure 3.

5 Methods

To demonstrate the utility of our annotated corpus, we explore two machine learning approaches for extracting actions and entities: a maximum entropy model and a neural network tagging model. We also present experiments for relation classification. We use the standard precision, recall and F_1 metrics to evaluate and compare the performance.

5.1 Maximum Entropy (MaxEnt) Tagger

In the maximum entropy model for action and entity extraction (Borthwick and Grishman, 1999), we used three types of features based on the current word and context words within a window of size 2:

- **Parts of speech features** which were generated by the GENIA POS Tagger (Tsuruoka and Tsujii, 2005), which is specifically tuned for biomedical texts;
- **Lexical features** which include unigrams, bigrams as well as their lemmas and synonyms from WordNet (Miller, 1995) are used;
- **Dependency parse features** which include dependent and governor words as well as the dependency type to capture syntactic information related to actions, entities and their contexts. We used the Stanford dependency parser (Chen and Manning, 2014).

5.2 Neural Sequence Tagger

We utilized the state-of-the-art Bidirectional LSTM with a Conditional Random Fields (CRF) layer (Ma and Hovy, 2016; Lample et al., 2016; Plank et al., 2016), initialized with 200-dimentional word vectors pretrained on 5.5 billion words from PubMed and PMC biomedical texts (Moen and Ananiadou, 2013). Words unseen in the pretrained vocabulary were randomly initialized using a uniform distribution in the range (-0.01, 0.01). We used Adadelta (Zeiler, 2012) optimization with a mini-batch of 16 sentences and trained each network with 5 different random seeds, in order to avoid any outlier results due to randomness in the model initialization.

5.3 Relation Classification

To demonstrate the utility of the relation annotations, we also experimented with a maximum entropy model for relation classification using features shown to be effective in prior work (Li and Ji, 2014; GuoDong et al., 2005; Kambhatla, 2004). The features are divided into five groups:

- **Word features** which include the words contained in both arguments, all words in between, and context words surrounding the arguments;
- **Entity type features** which include action and entity types associated with both arguments;
- **Overlapping features** which are the number of words, as well as actions or entities, in between the candidate entity pair;
- **Chunk features** which are the chunk tags of both arguments predicted by the GENIA tagger;
- **Dependency features** which are context words related to the arguments in the dependency tree according to the Stanford Dependency Parser.

Entity/Action (freq. in test set)	**MaxEnt**	**BiLSTM**	**BiLSTM + CRF**
Action (3519)	83.87	85.95	86.89
Amount (886)	68.25	81.59	82.34
Conc. (273)	56.84	65.36	76.36
Device (408)	49.14	58.73	64.02
Gen.-Measure (91)	05.88	06.45	25.68
Location (1007)	61.07	69.57	73.53
Meas.-Type (50)	15.38	18.75	21.62
Mention (37)	43.37	52.31	57.97
Method (177)	37.97	30.60	38.21
Modifier (720)	50.86	56.90	59.34
Numerical (129)	39.70	47.84	49.80
Reagent (2486)	60.54	71.34	74.55
Seal (43)	49.52	54.05	66.67
Size (69)	19.35	24.82	26.92
Speed (200)	74.88	85.31	91.00
Temperature (469)	80.69	86.68	91.90
Time (708)	83.68	92.69	93.94
pH (21)	41.86	53.66	70.00
Macro-avg F1	49.23	58.81	64.44
Micro-avg F1	68.03	74.99	78.03

Table 4: F1 scores for segmenting and classifying entities and action triggers compared across the various models.

MaxEnt Model	Relations		
Features	P	R	F1
Words	66.16	46.84	54.85
+ Entity Type	78.93	72.75	75.72
+ Overlap	80.81	74.73	77.65
+ Base Phrase Chunking	81.04	76.52	78.71
+ Dependency Tree	80.98	77.04	78.96

Table 5: Precision, Recall and F1 (micro-average) of the maximum entropy model for relation classification, as each feature is added.

Also included are features indicating whether the two spans are in the same noun phrase, prepositional phrase, or verb phrase.

6 Results

The full annotated dataset of 622 protocols are randomly split into training, dev and test sets using a 6:2:2 ratio. The training set contains 374 protocols of 8207 sentences, development set contains

MaxEnt Model	Actions			Entities		
Features	P	R	F1	P	R	F1
POS	74.83	79.94	77.30*	26.66	27.93	28.77
uni/bigram	76.29	69.59	72.79	43.75	32.93	37.58
POS, uni/bigram	79.77	85.51	82.54	49.83	54.51	52.07
POS, uni/bigram, lem./syn.	80.10	85.56	82.74	49.79	54.54	52.06
POS, uni/bigram, lem./syn., dep.	**81.65**	**86.22**	**83.87**	**57.04**	**63.03**	**59.90***

Table 6: Performance of maximum entropy model with various features.*The POS features are especially useful for recognizing actions; dependency based features are more helpful for entities than actions.

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 7: 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).

123 protocols of 2736 sentences, and test set contains 125 protocols of 2736 sentences. We use the evaluation script from the CoNLL-03 shared task (Tjong Kim Sang and De Meulder, 2003), which requires exact matches of label spans and does not reward partial matches. During the data preprocessing, all digits were replaced by ‘0’.

6.1 Entity Identification and Classification

Table 4 shows the performance of various methods for entity tagging. We found that the BiLSTM-CRF model consistently outperforms other methods, achieving an overall F1 score of 86.89 at identifying action triggers and 72.61 at identifying and classifying entities.

Table 6 shows the system performance of the MaxEnt tagger using various features. Dependency based features have the highest impact on the detection of entities, as illustrated by the absolute drop of 7.84% in F-score when removed. Parts of speech features alone are the most effective in capturing action words. This is largely due to action words appearing as verbs or nouns in the majority of the sentences as shown in Table 7. We also notice that the GENIA POS tagger, which is trained on Wall Street Journal and biomedical abstracts in the GENIA and PennBioIE corpora, under-identifies verbs in wet lab protocols. We suspect this is due to fewer imperative sentences in the training data. We leave further investigation for future work, and hope the release of our dataset can help draw more attention to NLP research on instructional languages.

6.2 Relation Classification

Finally, precision and recall at relation extraction are presented in Table 5. We used gold action and entity segments for the purposes of this particular evaluation. We obtained the best performance when using all feature sets.

7 Conclusions

In this paper, we described our effort to annotate wet lab protocols with actions and their semantic arguments. We presented an annotation scheme that is both biologically and linguistically motivated and demonstrated that non-experts can effectively annotate lab protocols. Additionally, we empirically demonstrated the utility of our corpus for developing machine learning approaches to shallow semantic parsing of instructions. Our annotated corpus of protocols is available for use by the research community.

Acknowledgement

We would like to thank the annotators: Bethany Toma, Esko Kautto, Sanaya Shroff, Alex Jacobs, Berkay Kaplan, Colins Sullivan, Junfa Zhu, Neena Baliga and Vardaan Gangal. We would like to thank Marie-Catherine de Marneffe and anonymous reviewers for their feedback.

References

Vaishnavi Ananthanarayanan and William Thies. 2010. Biocoder: A programming language for standardizing and automating biology protocols. In *Journal of Biological Engineering*.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. In *Transactions of the Association for Computational Linguistics (TACL)*.

Maxwell Bates, Aaron J Berliner, Joe Lachoff, Paul R Jaschke, and Eli S Groban. 2016. Wet lab accelerator: a web-based application democratizing laboratory automation for synthetic biology. In *ACS synthetic biology*.

Mario Bollini, Stefanie Tellex, Tyler Thompson, Nicholas Roy, and Daniela Rus. 2013. Interpreting and executing recipes with a cooking robot. In *Proceedings of International Symposium on Experimental Robotics (ISER)*.

Andrew Borthwick and Ralph Grishman. 1999. A maximum entropy approach to named entity recognition. *Ph. D. Thesis, Dept. of Computer Science, New York University*.

Satchuthananthavale RK Branavan, Harr Chen, Luke S Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

SRK Branavan, Luke S Zettlemoyer, and Regina Barzilay. 2010. Reading between the lines: Learning to map high-level instructions to commands. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Dallas Card, Amber E Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

David L Chen and Raymond J Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI)*.

Dipanjan Das, Desai Chen, André FT Martins, Nathan Schneider, and Noah A Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. 2005. Exploring various knowledge in relation extraction. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*.

Jermsak Jermsurawong and Nizar Habash. 2015. Predicting the structure of cooking recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19:i180–i182.

Ross D King, Jem Rowland, Stephen G Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N Soldatova, Andrew Sparkes, Kenneth E Whelan, and Amanda Clare. 2009. The automation of science. *Science*, 324(5923):85–89.

Klaus Krippendorff. 2004. *Content analysis: An introduction to its methodology*. SAGE Publications.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

George A Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LBM2013)*.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *The 54th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Larisa N Soldatova, Daniel Nadis, Ross D King, Piyali S Basu, Emma Haddi, Véronique Baumlé, Nigel J Saunders, Wolfgang Marwan, and Brian B Rudkin. 2014. EXACT2: the semantics of biomedical protocols. *BMC bioinformatics*.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning (CoNLL)*.

Yoshimasa Tsuruoka and Jun'ichi Tsujii. 2005. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *Proceedings of the conference on human language technology and empirical methods in natural language processing (HLT-EMNLP)*.

Viktor Vasilev, Chenkai Liu, Traci Haddock, Swapnil Bhatia, Aaron Adler, Fusun Yaman, Jacob Beal, Jonathan Babb, Ron Weiss, Douglas Densmore, et al. 2011. A software stack for specification and robotic execution of protocols for synthetic biological engineering. In *3rd International Workshop on Bio-Design Automation (IWBDA)*.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *arXiv:1212.5701*.

A Annotation Guidelines

The wet lab protocol dataset annotation guidelines 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 sub-sections we summarize 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.

A.1 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 mentions in a sentence or a step are deliberate but short descriptions of a task tying together various entities in a meaningful way. Some examples of action words, (categorized using GENIA POS tagger), are present in Table 7 along with their frequencies.

A.2 Entities

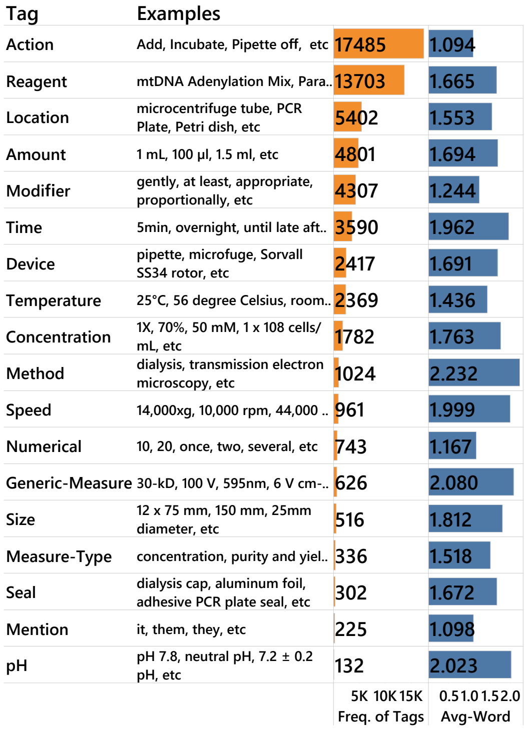
We broadly classify entities commonly seen in protocols under 17 tags. Each of the entity tags were designed to encourage short span length, with the average number of words per entity tag being 1.6. For example, `Concentration` tags are often very short: *60% 10x*, *10M*, *1 g/ml*, while the `Method` 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 4). 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.

A.2.1 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.

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



Tag	Examples	Freq. of Tags	Avg-Word
Action	Add, Incubate, Pipette off, etc	17485	1.094
Reagent	mtDNA Adenylation Mix, Para..	13703	1.665
Location	microcentrifuge tube, PCR Plate, Petri dish, etc	5402	1.553
Amount	1 mL, 100 µl, 1.5 ml, etc	4801	1.694
Modifier	gently, at least, appropriate, proportionally, etc	4307	1.244
Time	5min, overnight, until late aft..	3590	1.962
Device	pipette, microfuge, Sorvall SS34 rotor, etc	2417	1.691
Temperature	25°C, 56 degree Celsius, room..	2369	1.436
Concentration	1X, 70%, 50 mM, 1 x 108 cells/ mL, etc	1782	1.763
Method	dialysis, transmission electron microscopy, etc	1024	2.232
Speed	14,000xg, 10,000 rpm, 44,000 ..	961	1.999
Numerical	10, 20, once, two, several, etc	743	1.167
Generic-Measure	30-kD, 100 V, 595nm, 6 V cm-..	626	2.080
Size	12 x 75 mm, 150 mm, 25mm diameter, etc	516	1.812
Measure-Type	concentration, purity and yiel..	336	1.518
Seal	dialysis cap, aluminum foil, adhesive PCR plate seal, etc	302	1.672
Mention	it, them, they, etc	225	1.098
pH	pH 7.8, neutral pH, 7.2 ± 0.2 pH, etc	132	2.023

Figure 4: Examples, Frequency and Avg-Word for actions and entities.

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.

A.2.2 Measure Based Entities

Amount: The amount of any reagent being used in a 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: Duration of 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, but could also be a passive verb.

Speed: Typically a measure that represents rotation per min for centrifuges.

Numerical: A generic tag for a number that

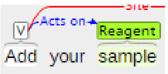
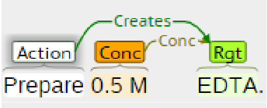
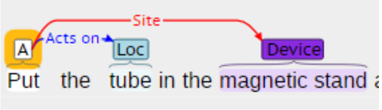
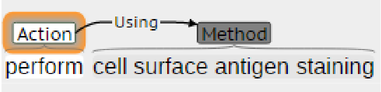
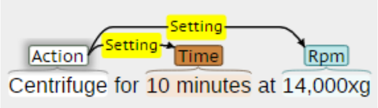
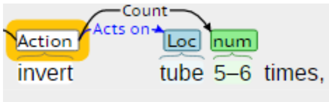
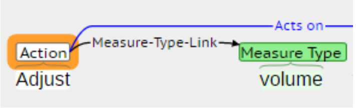
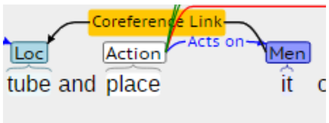
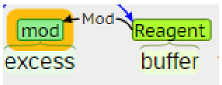
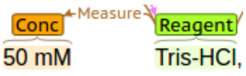
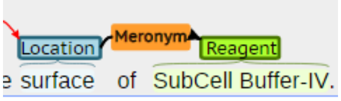
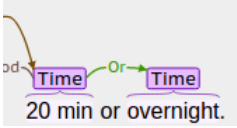
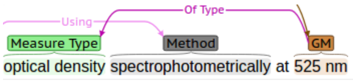
Label	Syntax/Rules	Example
Acts-on	Action ⇒ Reagent | Location | Mention | Device | Seal	
Creates	Action ⇒ Reagent | Mention	 Creates Conc Action Conc Rgt Prepare 0.5 M EDTA.
Site	Action ⇒ Location | Device | Mention | Reagent	 Site Acts on A Loc Device Put the tube in the magnetic stand
Using	Action ⇒ Method | Action | Seal | Device | Mention | Reagent | Location	 Using Action Method perform cell surface antigen staining
Setting	Action | Device | Modifier ⇒ Method | Action | Seal | Device | Mention | Reagent | Location	 Setting Setting Action Time Rpm Centrifuge for 10 minutes at 14,000xg
Count	Action ⇒ Numerical	 Count Acts on Action Loc num invert tube 5–6 times,
Measure-Type-Link	Action ⇒ Measure-Type	 Acts on Measure-Type-Link Action Measure Type Adjust volume
Coreference	Mention ⇒ [Every other entity]	 Coreference Link Acts on Loc Action Men tube and place it c
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	 Meronym Location Reagent e surface of SubCell Buffer-IV.
Or	[All Entities or Action] ⇒ [All Entities or Action]	
Of-Type	Generic-Measure | Numerical ⇒ Measure-Type	 Of Type Using Measure Type Method GM optical density spectrophotometrically at 525 nm

Table 8: Relations along with their rules and examples

doesn't fit time, temp, etc and which isn't accompanied by its unit of measure.

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

Size A measure of the dimension of an object. For example: length, area or thickness.

Measure-Type: A generic tag to mark the type of measurement associated with a number.

pH: measure of acidity or alkalinity of a solution.

A.2.3 Parts of Speech based Entities

Modifier: A word or a phrase that acts as an additional description of the entity it is modifying. For example, *quickly mix* vs *slowly mix* are clearly two different actions, informed by their modifiers "quickly" or "slowly" respectively.

Mention: Words that can refer to an object mentioned earlier in the sentence.

A.3 Relations

A.3.1 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 indicates that the Device or Location is the site where the action is performed. It is also used as a way to indicate which entity will finally hold/contain the result of the action.

Using: Any entity that the action verb makes use of is linked with this relation.

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 should take place.

Measure Type Link: Associates an action to a Measure Type entity that the Action is instructing to measure.

A.3.2 Binary Relations (Entity - Entity)

Coreference: A link that associates two phrases when those two phrases 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 associated with those settings.

Measure: A link that associates the various numerical measures to the entity its trying to measure directly.

Meronym: Links reagents, locations or devices with materials contained in the reagent, location or device.

Or: Allows chaining 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.