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

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

Abstract

We describe an effort to annotate a corpus of natural language instructions consisting of 662 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 will make our annotated corpus available to the research community upon publication.

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[1] and in the long run, enable robotic automation[2] by mapping natural language instructions to executable actions. In this study we take a first step towards this goal by (1) annotating a database of wet lab protocols with semantic actions and their arguments; and (2) 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 and large enough to enable machine learning approaches.

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.

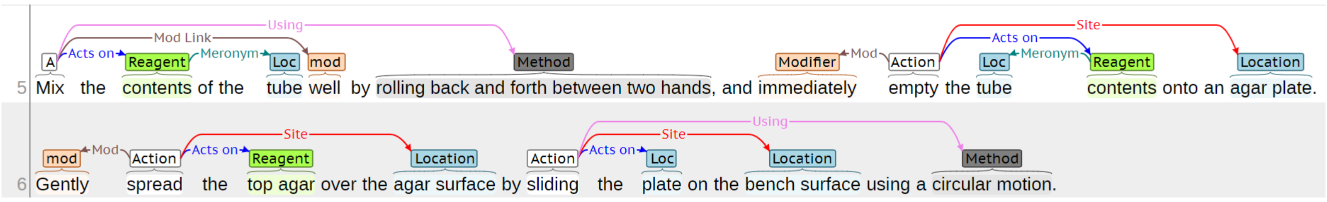
Figure 1: An example wet lab protocol. The first seven steps are imperative sentences, and the last sentence describes the end results and their subsequent utilization.

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 corpus. 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] http://klavinslab.org/aquarium-about.html

[2] https://github.com/Autodesk/bionano-wetLabAccelerator



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.

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. 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/stem cell biology, etc. 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 2); (3) The resulting corpus should be useful for training machine learning models to automatically extract experimental actions for downstream applications (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 describing the type of relations, whereas we defined and annotated 12 types of relations between actions and entities, or pairs of entities (see Appendix in the supplementary materials 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. Entity tags were designed to keep span length short, with the average number of words per mention being 1.6. 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.

	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 1: 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 2: Inter-annotator agreement (Krippendorff's α) between annotators with biology, linguistics and other backgrounds.

4 Annotation Process

Our final corpus consists of 622 protocols annotated by a team of 10 annotators. Corpus statistics are provided in Table 1. 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 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.

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-score 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 the contextual words within a window of size 2: (1) Parts of speech features which were generated by the GENIA POS Tagger (Tsuruoka and Tsujii, 2005) specifically tuned for biomedical texts; (2) Lexical features which include unigrams, bigrams as well as their lemmas and synonyms from WordNet (Miller, 1995) are used. (3) Dependency parse features which include dependent and governor words as well as the dependency type to capture syntactic information of actions /entities and their contexts. We used Stanford dependency parser (Chen and Manning, 2014). System performance using various features is presented in Table 3.

5.2 Neural Sequence Tagging

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), and initialized with 200-dimentional word vectors pretrained on 5.5 billion words from PubMed and PMC biomedical texts from (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:

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 3: 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.

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
Concentration (273)	56.84	65.36	76.36
Device (408)	49.14	58.73	64.02
Generic-Measure (91)	05.88	06.45	25.68
Location (1007)	61.07	69.57	73.53
Measure-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
Full Macro-avg F1	49.23	58.81	64.44
Full Micro-avg F1	68.03	74.99	78.03

Table 4: F1 scores for segmenting 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 at classifying relations as each feature set is added.

(1) Words features, which include the words contained in both arguments, all words in between, and context words surrounding the arguments; (2) Entity type features, which include action and entity types associated with both arguments; (3) Overlapping features, which are the number of words, as well as actions or entities, in between the candidate entity pair; (4) Chunk features, that are the chunk tags of both arguments predicted by the GENIA tagger; (5) Dependency features, which are context words related to the arguments in the dependency tree according to the Stanford Dependency Parser. Also included are features indicating whether the two spans are in the same noun phrase, preposition phrase, or verb phrase.

6 Results

The full annotated dataset of 622 protocols are randomly split into training, dev and test sets by 6:2:2 ratio. The training set contains 374 protocols of 8207 sentences, development set contains 123 protocols of 2736 sentences, and test set contains 125 protocols of 2736 sentences. We use the evaluation script from 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 of the digits were replaced by ‘0’. Performance of the various methods is presented in Table 4. We found that the BiLSTM-CRF consistently outperforms other methods, achieving an overall F1 score of 86.89 at identifying action triggers and 72.61 at identifying and classifying entities. 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 comparing to biomedical and linguistic experts. Additionally, we empirically demonstrated the utility of our corpus for developing machine learning approaches to shallow semantic parsing of instructions. We will make our annotated protocols available to the research community.[3]

[3] Our annotated dataset will be available at: https://anonymized_url/ (this is a preprint).

Acknowledgement

We would like to thank Esko Kautto and Sanaya Shroff for their biomedical expertise and Bethany Toma for her linguistic expertise in curating and validating the annotation guidelines. We would also like to thank them for their annotation effort along with all the annotators Alex Jacobs, Berkay Kaplan, Colins Sullivan, Junfa Zhu, Neena Baliga and Vardaan Gangal for their meticulous effort in creating a high quality annotated dataset.

References

Vaishnavi Ananthanarayanan and William Thies. 2010. Biocoder: A programming language for standardizing and automating biology protocols. *Journal of biological engineering*, 4(1):13.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62.

Mario Bollini, Stefanie Tellex, Tyler Thompson, Nicholas Roy, and Daniela Rus. 2013. Interpreting and executing recipes with a cooking robot. In *Experimental Robotics*, pages 481–495. Springer.

Andrew Borthwick and Ralph Grishman. 1999. *A maximum entropy approach to named entity recognition*. Ph.D. thesis, New York University, Graduate School of Arts and Science.

Satchuthananthavale RK Branavan, Harr Chen, Luke S Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 82–90. Association for Computational Linguistics.

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*, pages 1268–1277. Association for Computational Linguistics.

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 *ACL*.

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)*, pages 740–750.

David L Chen and Raymond J Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *AAAI*, volume 2, pages 1–2.

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*, pages 427–434. Association for Computational Linguistics.

Jermsak Jermsurawong and Nizar Habash. 2015. Predicting the structure of cooking recipes. In *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*, page 22. Association for Computational Linguistics.

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

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

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *ACL (1)*, pages 402–412.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

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.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. *arXiv preprint arXiv:1604.05529*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *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*, 15(14):S5.

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 at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

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*, pages 467–474. Association for Computational Linguistics.

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*.

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