

An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols Chaitanya Kulkarni, Wei Xu, Alan Ritter, Raghu Machiraju The Ohio State University

Introduction

Cumbersome biological experiments necessitates automation to reduce human error and make science reproducible

Isolation of temperate phages by plaque agar overlay **1.** Melt soft agar overlay tubes in boiling water and place in the 47° C water bath.

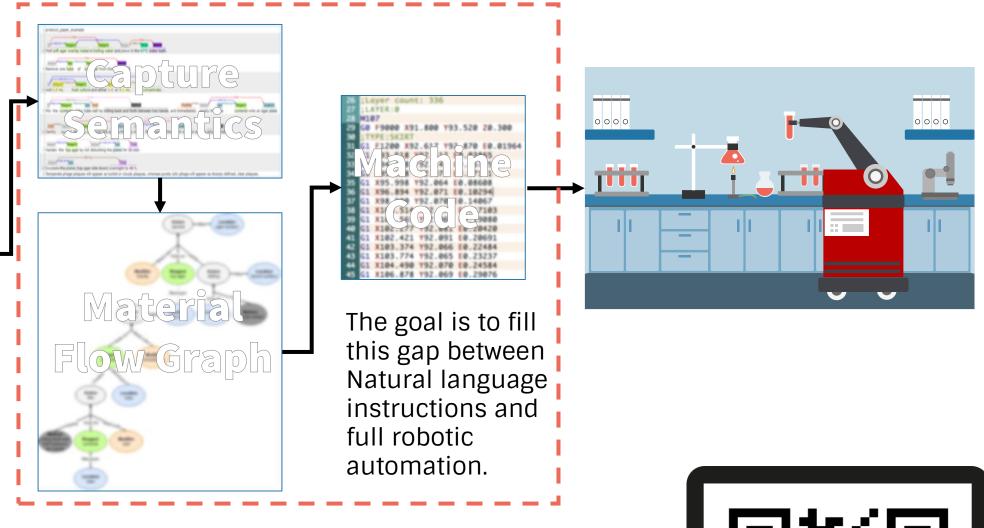
2. Remove one tube of soft agar from the water bath. **3.** Add 1.0 mL host culture and either 1.0 or 0.1 mL viral

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

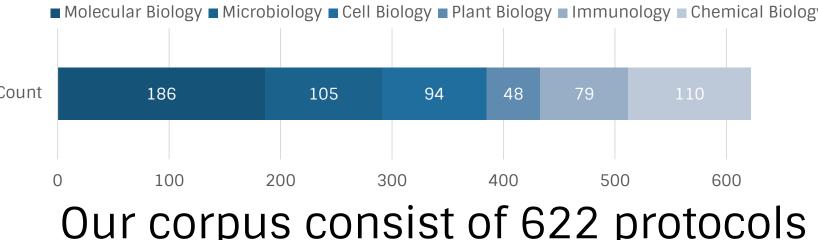
7. Harden the top agar by not disturbing the plates for 30

8. Incubate the plates (top agar side down) overnight to



- We take the first steps towards this goal by, Introducing canonical semantic representations understood by experts and non-experts and create a comprehensive corpora, WLP
- Demonstrating utility of corpora by developing machine learning approaches for semantic parsing of instructions

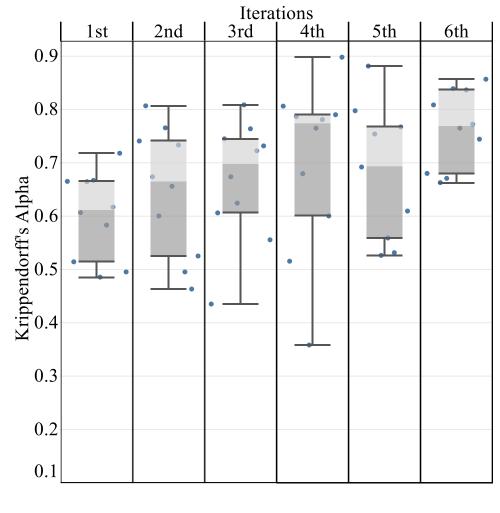
Corpus Statistics

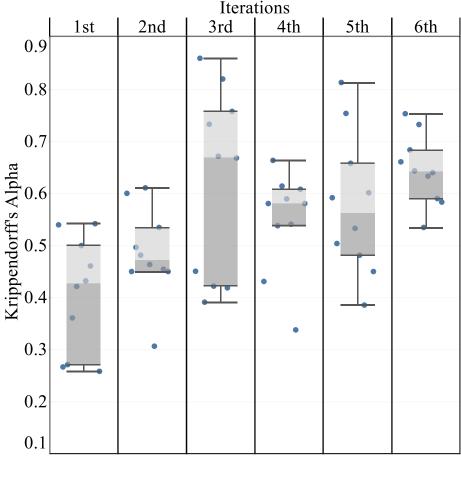


	Total	per Protocol	per Sentence
# of sentences	13679	21.99	_
# of words	177770	285.80	12.99
# of entities	43236	69.51	3.16
# of relations	42425	68.21	3.10
# of actions	17485	28.11	1.28

Our corpus consist of 622 protocols annotated by a team of 10 annotators

Inter-Annotator Agreement





Every iteration consisted of 10 protocols, annotated by 4 coders.

Inter-annotator agreement improves over iterations as changes were made in the annotation guidelines.

Action + Entities

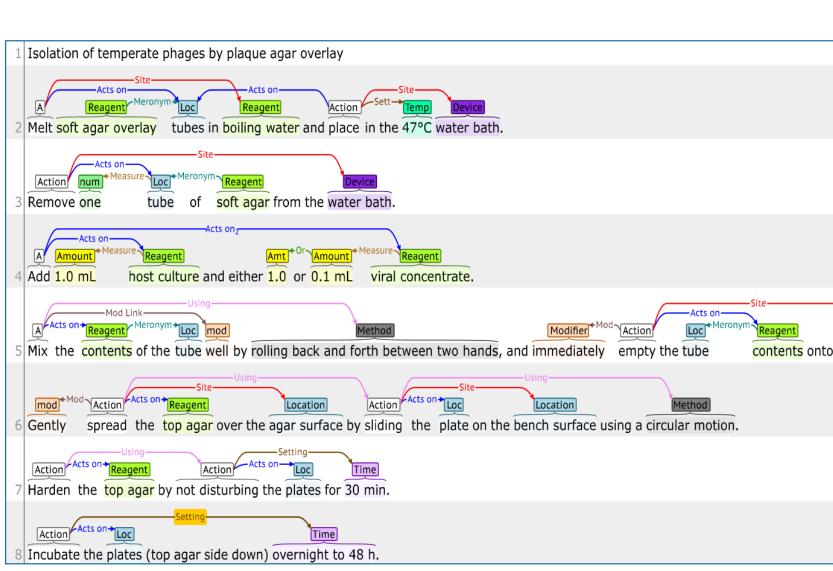
Relations

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

Similar inter annotator agreement between annotators with varying backgrounds demonstrates the ease of comprehension from experts and nonexperts alike.

Wet Lab Protocol Corpus





WLP corpus constructs canonical representation for a wet lab protocol an action graph directly derived from the annotations.

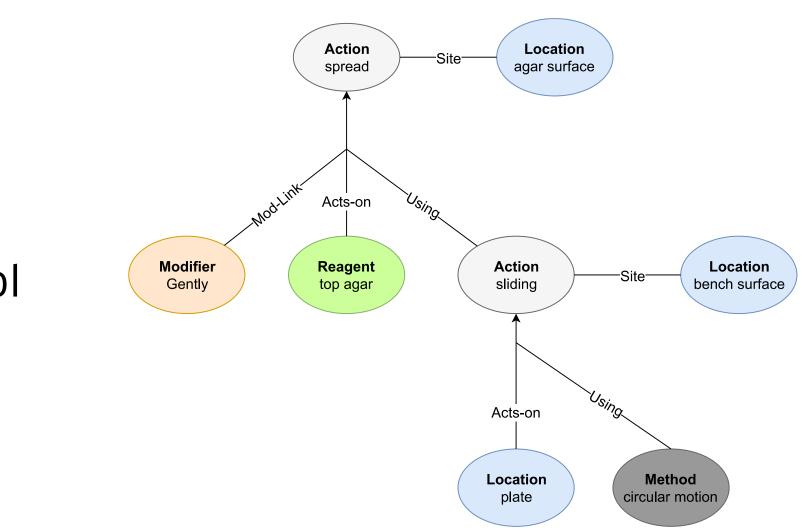
Entity Extraction

Тад	Examples		MaxEnt
Reagent	mtDNA Adenylation Mix, Paraformaldehyde, etc	13703	1.665
Location	microcentrifuge tube, PCR Plate, Petri dish, etc	5402	1.553
Amount		4801	1.694
Modifier	gently, at least, appropriate, proportionally, etc	<mark>4307</mark>	1.244
Time	5min, overnight, until late afternoon, etc	<mark>359</mark> 0	1.962
Device	pipette, microfuge, Sorvall SS34 rotor, etc	<mark>24</mark> 17	1.691
Temperature	25°C, 56 degree Celsius, room temperature, etc	<mark>23</mark> 69	1.436
Concentration	1X, 70%, 50 mM, 1 x 108 cells/ mL, etc	1782	1.763
Method	dialysis, transmission electron microscopy, etc	<mark>1</mark> 024	2.232
Speed	14,000xg, 10,000 rpm, 44,000 rcf, etc	961	1.999
Numerical	10, 20, once, two, several, etc	743	1.167
Generic-Meas	30-kD, 100 V, 595nm, 6 V cm-1, 140 bp, etc	626	2.080
Size	12 x 75 mm, 150 mm, 25mm diameter, etc	516	1.812
Measure-Type	concentration, purity and yield, absorbance, etc	336	1.518
Seal	dialysis cap, aluminum foil, adhesive PCR plate seal, etc	302	1.672
Mention	it, them, they, etc	225	1.098
рН	pH 7.8, neutral pH, 7.2 ± 0.2 pH, etc	132	2.023
		OK 5K 10K Freq. of Tags	0.5 1.0 Avg Wo

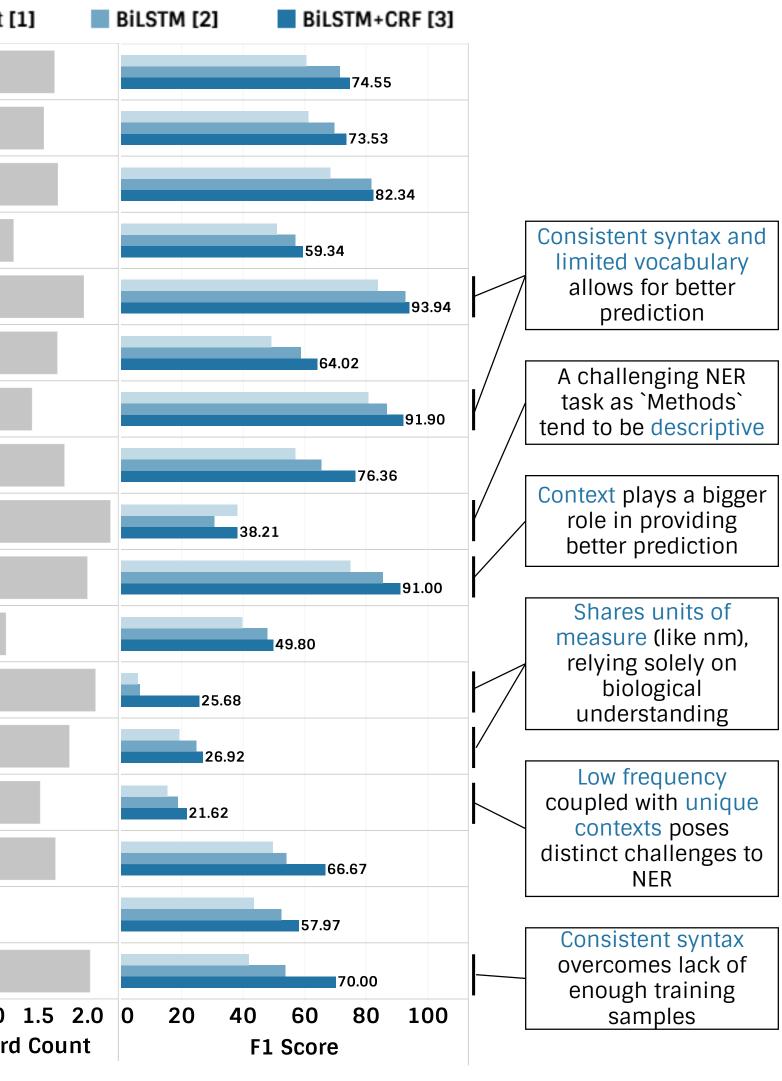
- York University
- IEEE International Conference on Acoustics, Speech and Signal Processing, pages 6645–6649. Association for Computational Linguistics (ACL).

Wet lab protocols are sequence of steps consisting

- Imperative Sentences: instructing actions
- **Declarative Sentences**: describing result of a previous action
- **Notes**: general guidelines and/or warnings

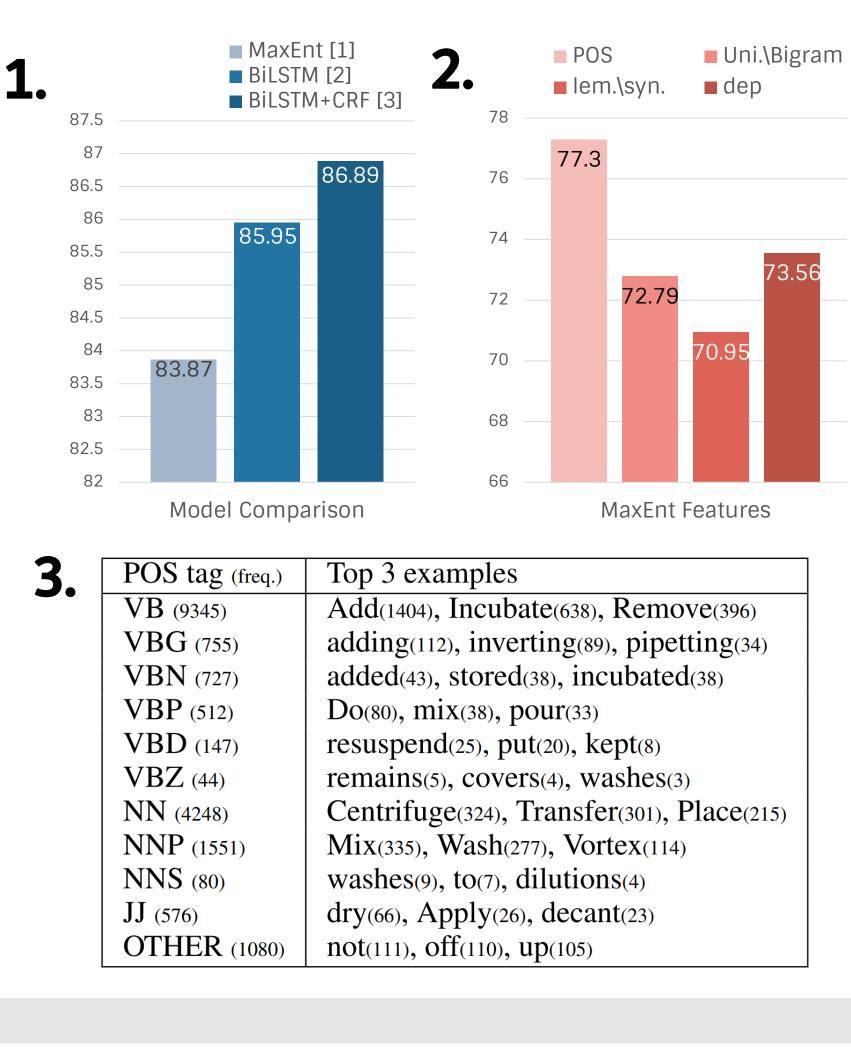






References

Andrew Borthwick and Ralph Grishman. 1999. A maximum entropy approach to named entity recognition. *Ph. D. Thesis, Dept. of Computer Science, New* Alan Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 Xuezhe Ma and Eduard Hovy. 2016. End-to-end se-quence labeling via bi-directional lstm-cnns-crf. In Proceedings of the 54th Annual Meeting of the



Relation	Examples
Types	•
Acts-on	Add your sample
Setting	Action Setting Time Centrifuge for 10 minute
Measure	Conc ^{Measure} Reagent 50 mM Tris-HCl,
Site	Acts on Loc Put the tube in the magn
Using	Action Using Me
Mod-Link	excess buffer
Meronym	Location Meronym Reager surface of SubCell Bu
Or	Time Or Time 20 min or overnight.
Creates	Action Creates Rgt Prepare 0.5M EDTA.
Count	Action Acts on Loc num invert tube 5–6 time
Measure- Type-Link	Action Measure-Type-Link- Adjust the
Coreference	Loc Action Acts on M tube and place
Of-Type	Action Measure Type action optical density spectrop

Conclusion + Future Work

We present a corpus with accessible semantic representation. Given the varying emphasis on morphology and context, every named entity and relation in this representation poses unique challenges to semantic parsing.

In addition to implementing methods that address these challenges, we plan to extend the corpus by inter connecting all the sentences to build a more complete protocol representation.

Action Extraction

1. Evaluate corpora by classifying actions using the best maximum entropy model and 2 neural models.

2. Parts of speech was the most effective in capturing action words

3. Majority of the action verbs fall under VBs (60.48%) or NN (30.84%) using GENIA POS tagger. A small percentage are misclassified under **OTHER** (5.66%) and JJ (3.02%).

