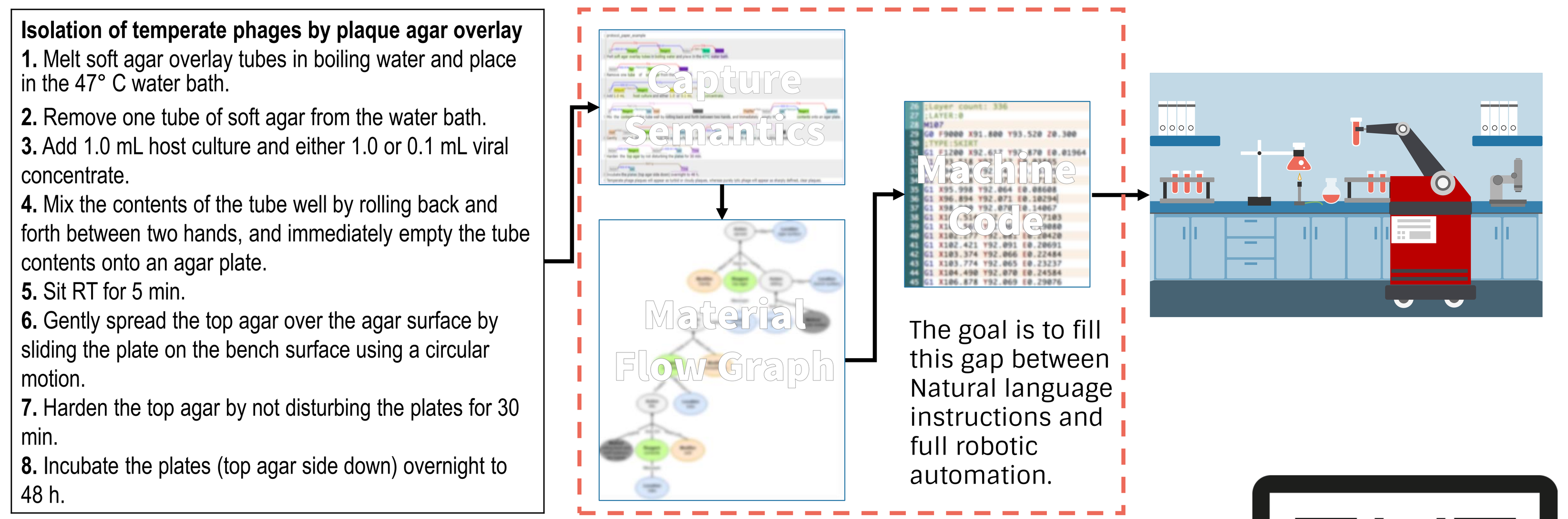
An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols

Chaitanya Kulkarni, Wei Xu, Alan Ritter, Raghu Machiraju
The Ohio State University

THE OHIO STATE UNIVERSITY

Introduction

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

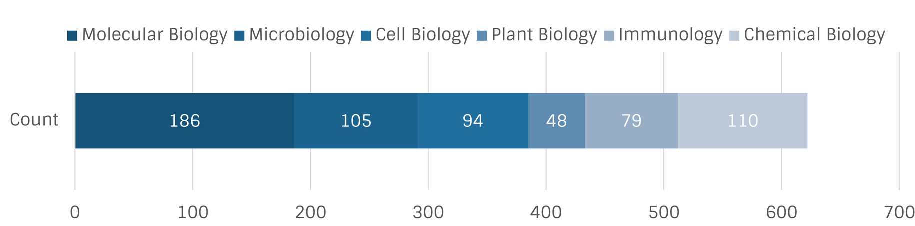


We take the first steps towards this goal by,

- 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

Web Demo

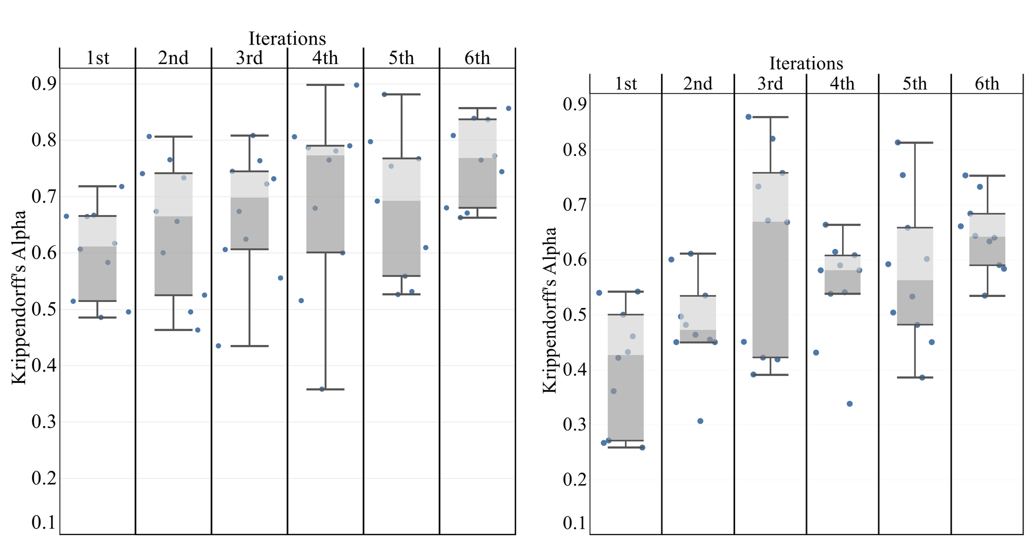
Corpus Statistics



	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



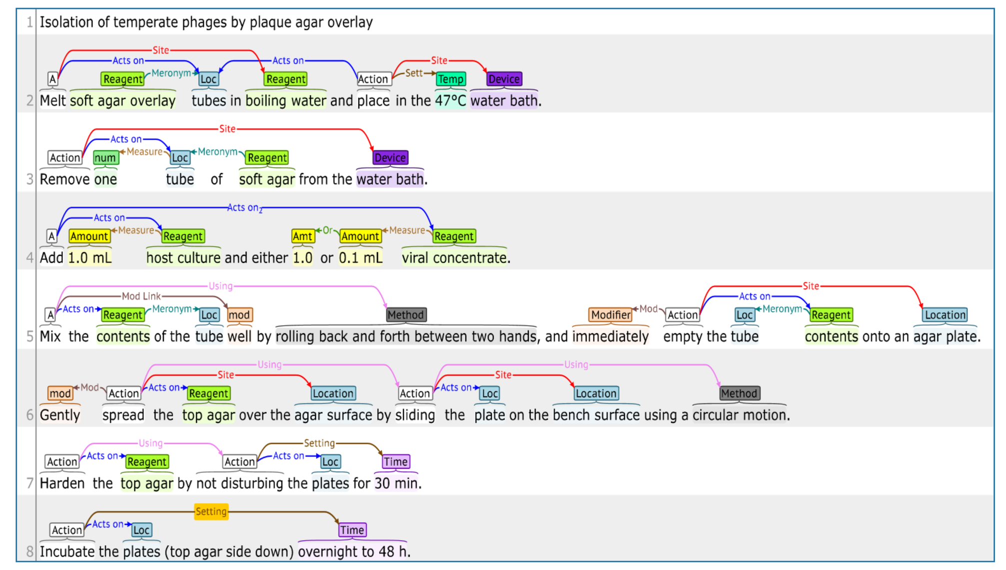
Every iteration consisted of 10 protocols, annotated by 4 coders.

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

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 non-experts alike.

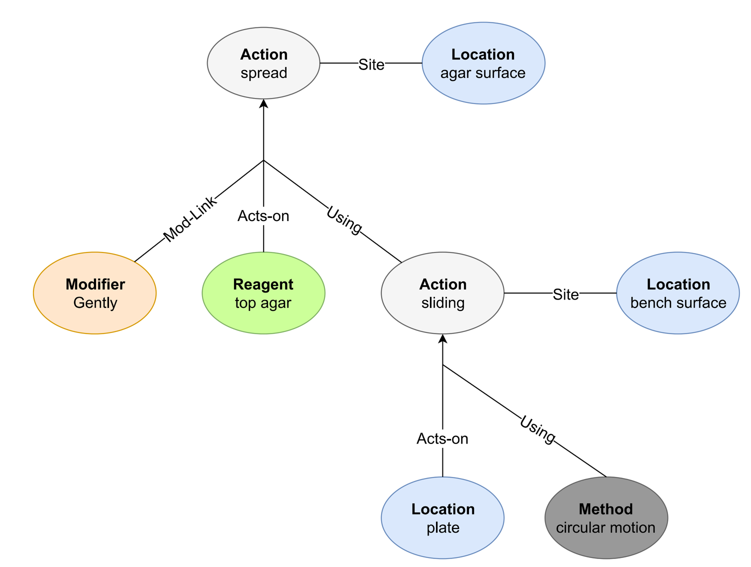
Wet Lab Protocol Corpus



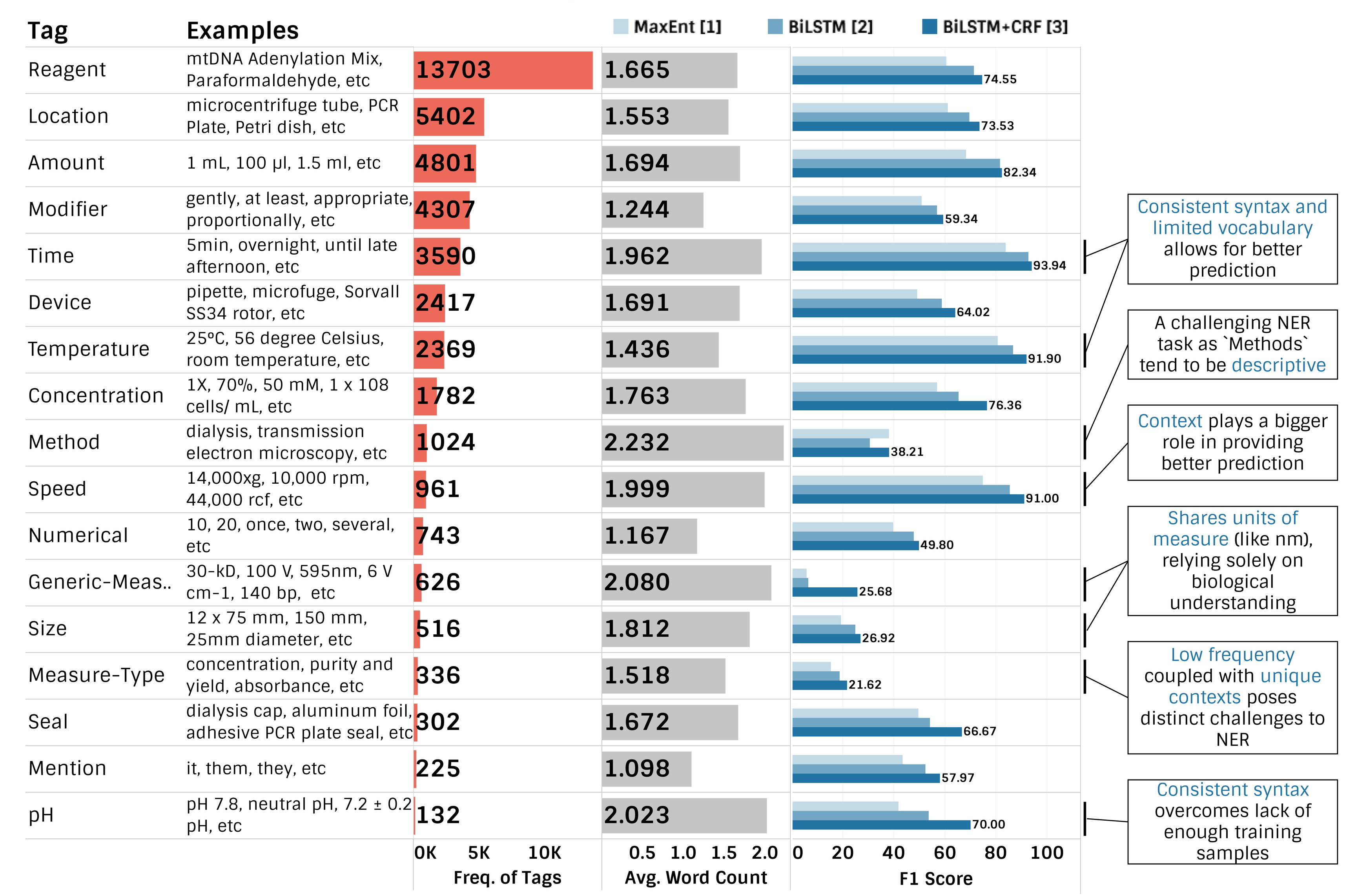
Wet lab protocols are sequence of steps consisting of:

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

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



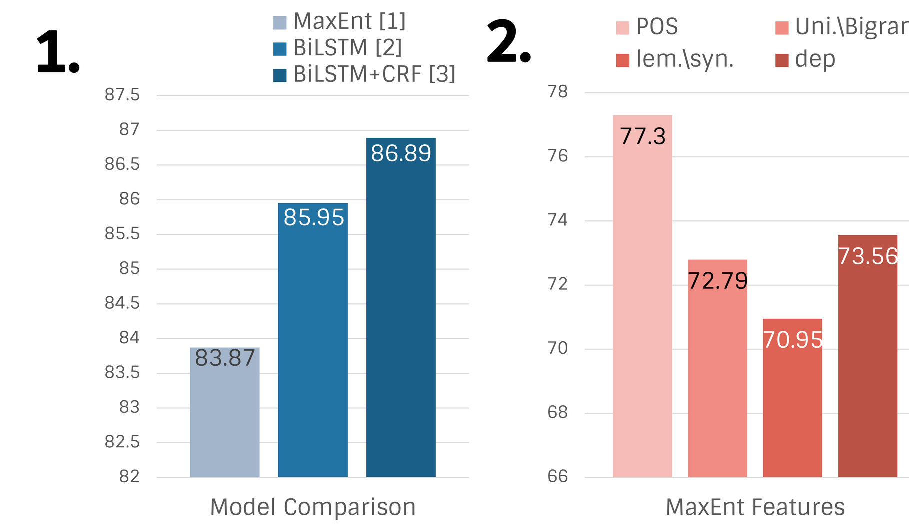
Entity Extraction



References

1. Andrew Borthwick and Ralph Grishman. 1999. A maximum entropy approach to named entity recognition. *Ph. D. Thesis, Dept. of Computer Science, New York University.*
2. Alan Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. *In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649.
3. 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 Association for Computational Linguistics (ACL).*

Action Extraction



3.

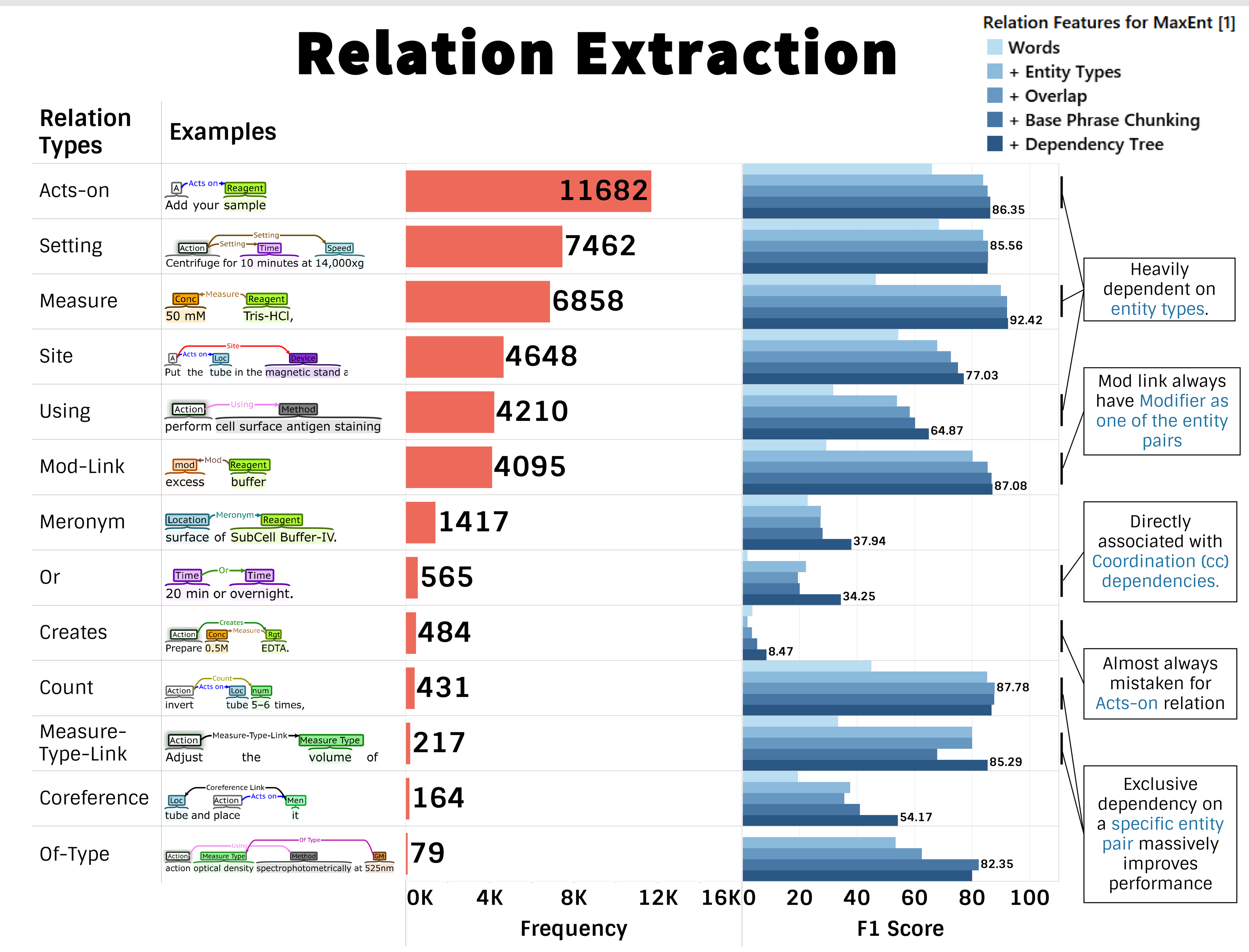
POS tag (freq.)	Top 3 examples
VB (9345)	Add(1404), Incubate(638), Remove(396)
VBG (755)	adding(112), inverting(89), pipetting(34)
VBN (727)	added(43), stored(38), incubated(38)
VBP (512)	Do(80), mix(38), pour(33)
VBD (147)	resuspend(25), put(20), kept(8)
VBZ (44)	remains(5), covers(4), washes(3)
NN (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)

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%).

Relation Extraction



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.

kulkarni.132@osu.edu