WNUT-2020 Task 1: Extracting Entities and Relations from Wet Lab Protocols

Jeniya Tabassum[1], Wei Xu[2], Alan Ritter[2]
[1] Department of Computer Science and Engineering, The Ohio State University
[2] School of Interactive Computing, Georgia Institute of Technology
bintejafar.1@osu.edu {wei.xu, alan.ritter}@cc.gatech.edu

Abstract

This paper presents the results of the wet lab information extraction task at WNUT 2020. This task consisted of two sub tasks- (1) a named entity recognition task with 13 participants; and (2) a relation extraction task with 2 participants. We outline the task, data annotation process, corpus statistics, and provide a high-level overview of the participating systems for each sub task.

1 Introduction

Wet Lab protocols consist of natural language instructions for carrying out chemistry or biology experiments (for an example, see Figure 1). While there have been efforts to develop domain-specific formal languages in order to support robotic automation[1] of experimental procedures (Bates et al., 2017), the vast majority of knowledge about how to carry out biological experiments or chemical synthesis procedures is only documented in natural language texts, including in scientific papers, electronic lab notebooks, and so on.

Recent research has begun to apply human language technologies to extract structured representations of procedures from natural language protocols (Kuniyoshi et al., 2020; Vaucher et al., 2020; Kulkarni et al., 2018; Soldatova et al., 2014; Vasilev et al., 2011; Ananthanarayanan and Thies, 2010). Extraction of named entities and relations from these protocols is an important first step towards machine reading systems that can interpret the meaning of these noisy human generated instructions.

However, performance of state-of-the-art tools for extracting named entity and relations from wet lab protocols still lags behind well edited text genres (Jiang et al., 2020). This motivates the need for continued research, in addition to new datasets and tools adapted to this noisy text genre.

[1] https://autoprotocol.org/

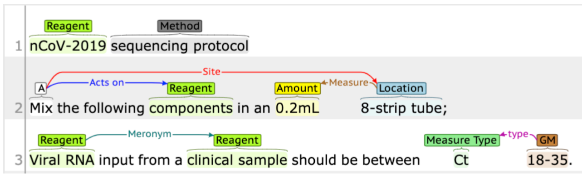

Figure 1: Examples of named entities and relations in a wet lab protocol

In this overview paper, we describe the development and findings of a shared task on named entity and relation extraction from the noisy wet lab protocols, which was held at the 6-th Workshop on Noisy User-generated Text (WNUT 2020) and attracted 15 participating teams.

In the following sections, we describe details of the task including training and development datasets in addition to the newly annotated test data. We briefly summarize the systems developed by selected teams, and conclude with results.

2 Wet Lab Protocols

Wet lab protocols consist of the guidelines from different lab procedures which involve chemicals, drugs, or other materials in liquid solutions or volatile phases. The protocols contain a sequence of steps that are followed to perform a desired task. These protocols also include general guidelines or warnings about the materials being used. The publicly available archive of protocol.io contains such guidelines of wet lab experiments, written by researchers and lab technicians around the world. This protocol archive covers a large spectrum of experimental procedures including neurology, epigenetics, metabolomics, stem cell biology, etc. Figure 1 shows a representative wet lab protocol.

The wet lab protocols, written by users from all over the worlds, contain domain specific jargon as well as numerous nonstandard spellings, abbreviations, unreliable capitalization. Such diverse and

	Train	Dev	Test-18	Test-20	Total
#protocols	370	122	123	111	726
#sentences	8444	2839	2813	3562	17658
#tokens	107038	36106	36597	51688	231429
#entities	48197	15972	16490	104654	185313
#relations	32158	10812	11242	70591	124803

	per Protocol	per Sentence
avg. #sentences	24.32	-
avg. #tokens	318.77	13.11
avg. #entities	255.25	10.49
avg. #relation	171.90	7.07

Table 1: Statistics of the Wet Lab Protocol corpus.

noisy style of user created protocols imposed crucial challenges for the entity and relation extraction systems. Hence, off-the-shelf named entity recognition and relation extraction tools, tuned for well edited texts, suffer a severe performance degradation when applied to noisy protocol texts (Kulkarni et al., 2018).

To address these challenges, there has been an increasing body of work on adapting entity and relation extraction recognition tools for noisy wet lab texts (Jiang et al., 2020; Luan et al., 2019; Kulkarni et al., 2018). However, different research groups have used different evaluation setups (e.g., training / test splits) making it challenging to perform direct comparisons across systems. By organizing a shared evaluation, we hope to help establish a common evaluation methodology (for at least one dataset) and also promote research and development of NLP tools for user generated wet-lab text genres.

2.1 Train and Development data

The training and development dataset for our task was taken from previous work on wet lab protocol (Kulkarni et al., 2018), which identifies 20 entities and 31 relations from the 623 protocols. We excluded the eight duplicate protocols from this dataset and then re-annotated the 615 unique protocols in BRAT (Stenetorp et al., 2012). This re-annotation process aided us to add the previously missing 20,613 missing entities along with 10,824 previously missing relations and also to facilitate removing the inconsistent annotations. The updated corpus statics is provided in Table 1. This full dataset (Train, Dev, Test-18) was provided to the participants at the beginning of the task and they were allowed to use any of part of this dataset to train their final model.

2.2 Test Data Annotation

For this shared task we added 111 new protocols (Test-20) which were used to evaluate the submitted models. Test-20 dataset consists of 100 randomly sampled general protocols and 11 manually selected covid-related protocols from protocols.io. This 111 protocols were double annotated by three annotators using a web-based annotation tool, BRAT (Stenetorp et al., 2012). Figure 1 presents a screenshot of our annotation interface. We also provided the annotators a set of guidelines containing the entity and relation type definitions. The annotation task was split in multiple iterations. In each iteration, an annotator was given a set of 10 protocols. An adjudicator then went through all the entity and relation annotations in these protocols and resolved the disagreements. Before adjudication, the inter-annotator agreement is 0.75 , measured by Cohen's Kappa (Cohen, 1960).

2.3 Baseline Model

We provided the participants baseline model for both of the subtasks. The baseline model for named entity recognition task utilized a feature-based CRF tagger developed using the CRF-Suite[2] with a standard set of contextual, lexical and gazetteer features. The baseline relation extraction system employed a feature-based logistic regression model developed using the Scikit-Learn[3] with a standard set of contextual, lexical and gazetteer features.

2.4 NER Systems

Thirteen teams (Table 3) participated in the named entity recognition sub-task. A wide variety of approaches were taken to tackle this task. Table 2 summarizes the word representations, features and the machine learning approaches taken by each team. Majority of the teams (11 out of 13) utilized contextual word representations. Four teams combined the contextual word representations with global word vectors. Only two teams did not use any type of word representations and relied entirely on hand-engineered features and a CRF taggers. The best performing teams utilized a combination of contextual word representation with ensemble of learning. Below we provide a brief description of the approach taken by each team.

[2] http://www.chokkan.org/software/crfsuite/

[3] https://scikit-learn.org/

Team	Word Representation	Features	Approach
BiTeM	BERT, BioBERT, RoBERTa, XLNet	-	Ensemble of Transformers
PublishInCovid19	PubMedBERT	-	Ensemble of BiLSTM-CRFs
Fancy Man	BERT	-	BERT fine tuning
mahab	BERT	Lexical Features	BERT fine tuning
mgsohrab	SciBERT	Lexical Features	SciBERT fine tuning
SudeshnaTCS	XLNet	Rules	XLNet fine tuning
Winners	BioBERT	-	BioBERT fine tuning
B-NLP	SciBERT, word2vec	-	Biaffine Classifier
BIO-BIO	BioBERT	-	BiLSTM-CRF
DSC-IITISM	GLoVe, CamemBERT, Flair	-	BiLSTM-CRF
Kabir	GLoVe, ELMo, BERT, Flair	Gazetteers	RNN-CRF
IBS	-	Gazetteers, POS Tagger	Ensemble of CRFs
KaushikAcharya	-	POS Tagger, Dependency Parser	CRF

Table 2: Summary of NER systems designed by each team.

Team Name	Affiliation
B-NLP	Bosch Center for Artificial Intelligence
Big Green	Dartmouth College
BIO-BIO	Harbin Institute of technology, Shenzhen
BiTeM	University of Applied Sciences and Arts of Western Switzerland, Swiss Institute of Bioinformatics, University of Geneva
DSC-IITISM	IIT(ISM) Dhanbad
Fancy Man	University of Manchester, Xian Jiaotong University, East China University of Science and Technology, Zhejiang University
IBS	IBS Software Pvt. Ltd, NTNU
Kabir	Microsoft
KaushikAcharya	Philips
mahab	Amirkabir University of Technology
mgsohrab	National Institute of Advanced Industrial Science and Technology
PublishInCovid19	Flipkart Private Limited
SudeshnaTCS	TCS Research & Innovation Lab
Winners	IIT, Kharagpur

Table 3: Team Name and affiliation of the participant.

B-NLP (Lange et al., 2020) modeled the NER as a parsing task and uses a biaffine classifier. The second classifier of their system used the predictions from the first classifier and then updates the labels of the predicted entities. Both of the classifiers utilized word2vec (Mikolov et al., 2013) and SciBERT (Lee et al., 2019) word representations.

BIO-BIO (Kecheng et al., 2020) implemented a BiLSTM-CRF tagger that utilized BioBERT (Lee et al., 2020) word representation.

BiTeM (Knafou et al., 2020) developed a voting based ensemble classifier that consists of 14 transformer models that utilized 7 different word representations including BERT (Devlin et al., 2019), ClinicalBERT (Huang et al., 2019), PubMedBERT(base) (Gu et al., 2020), BioBERT (Lee et al., 2020), RoBERTa (Liu et al., 2019), BiomedRoBERTa(base) (Gururangan et al., 2020) and XLNet (Yang et al., 2019).

DSC-IITISM (Gupta et al., 2020) developed a BiLSTM-CRF model that utilized a concatenation of CamemBERT(base) (Martin et al., 2020), Flair(PubMed) (Akbik et al., 2018), and GloVe(en) (Pennington et al., 2014) word representations.

Fancy Man (Zeng et al., 2020) fine-tuned the BERT(base) (Devlin et al., 2019) model with an additional linear layer.

IBS (Sikdar et al., 2020) system used an ensemble classifier consists of 4 feature based on CRF taggers.

Kabir (Khan, 2020) system used a RNN-CRF model that utilized a concatenation of Flair(PubMed) (Akbik et al., 2018) and ELMo(PubMed) (Peters et al., 2018) word representation.

KaushikAcharya (Acharya, 2020) system used a linear CRF with hand-crafted features.

mahab (Pour and Farinnia, 2020) system fine-tuned the BERT(base) (Devlin et al., 2019) sequence tagging model.

mgsohrab (Sohrab et al., 2020) fine-tuned the SciBERT (Beltagy et al., 2019) model.

PublishInCovid19 (Singh and Wadhawan, 2020) used a structured ensemble classifier (Nguyen and Guo, 2007) consisting of 11 BiLSTM-CRF taggers, that utilized the PubMedBERT (Gu et al., 2020) word representation.

SudeshnaTCS (Jana, 2020) fine-tuned XLNet (Yang et al., 2019) model.

	P	R	F_1
Exact Match			
BiTeM	**84.73**	72.25	**77.99**
PublishInCovid19	81.36	**74.12**	77.57
Fancy Man	76.21	71.76	73.92
mahab	50.19	52.96	51.54
mgsohrab	83.69	70.62	76.60
SudeshnaTCS	74.99	71.43	73.16
Winners	77.00	72.93	74.91
B-NLP	77.95	63.93	70.25
BIO-BIO	78.49	71.06	74.59
DSC-IITISM	64.20	57.07	60.42
Kabir	78.79	72.20	75.35
IBS	74.26	62.55	67.90
KaushikAcharya	73.68	63.98	68.48
Partial Match			
BiTeM	**88.72**	75.66	81.67
PublishInCovid19	85.74	**78.11**	**81.75**
Fancy Man	81.15	76.41	78.71
mahab	55.09	58.14	56.57
mgsohrab	87.95	74.22	80.50
SudeshnaTCS	79.73	75.95	77.80
Winners	81.76	77.43	79.54
B-NLP	84.85	69.59	76.46
BIO-BIO	83.16	75.29	79.03
DSC-IITISM	68.52	60.90	64.49
Kabir	83.73	76.73	80.08
IBS	79.72	67.15	72.89
KaushikAcharya	79.31	68.87	73.73

Table 4: Results on extraction of 20 Named Entity types from the *Test-20* dataset. **Exact Match** reports the performance when the predicted entity type is same as the gold entity and the predicted entity boundary is the exact same as the gold entity boundary. **Partial Match** reports the performance when the predicted entity type is same as the gold entity and the predicted entity boundary has some overlap with gold entity boundary.

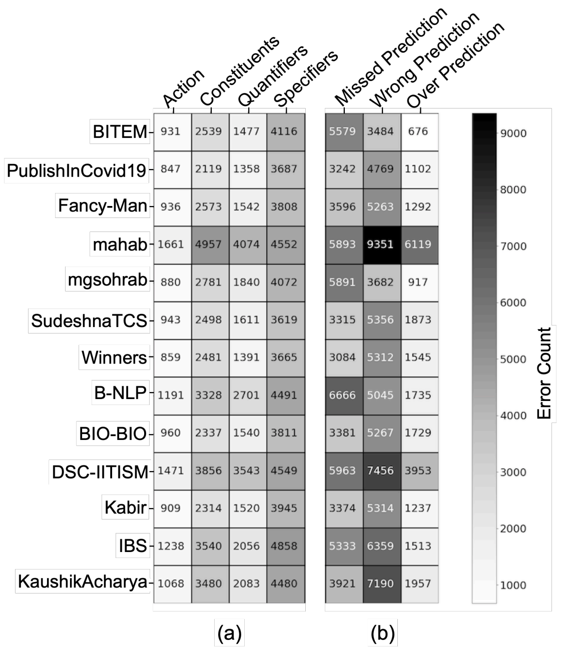

Figure 2: Summary of incorrectly classified entity tokens by each submitted systems.The 'Action' column represents the counts where the system failed to predict an action entity. The'Constituents' column represents the counts where the system failed to predict any of the constituent type entities: 'Reagent', 'Location', 'Device', 'Mention', and 'Seal' entity types. The 'Quantifiers' column represents the counts where the system failed to predict any of the quantifying entities: 'Amount', 'Concentration', 'Size', 'Time', 'Temperature', 'pH', 'Speed', 'Generic-Measure' and 'Numerical' entity types. The 'Specifiers' column represents the counts where the system failed to predict any of the constituent type entities: 'Modifier', 'Measure-Type' and 'Method' entity types.

Winners (Kaushal and Vaidhya, 2020) fine-tuned the Bio-BERT (Lee et al., 2020) model.

2.5 Relation Extraction Systems

Two teams (Table 3) participated in the relation extraction sub-task. Both of the teams followed fine-tuning of contextual word representation and did not use any hand-crafted features. Table 5 summarizes the word representations and the machine learning approaches followed by each team. Below we provide a brief description of the model developed by taken by each team.

Big Green (Miller and Vosoughi, 2020) considered the protocols as a knowledge graph, in which relationships between entities are edges in the knowledge graph. They trained a BERT (Devlin et al., 2019) based system to classify edge presence and type between two entities, given entity text, label, and local context.

mgsohrab (Sohrab et al., 2020) utilized PubMedBERT (Gu et al., 2020) as input to the relation extraction model that enumerates all possible pairs of arguments using deep exhaustive span representation approach.

3 Evaluation

In this section, we present the performance of each participating systems along with a description of the errors made by the model types.

3.1 NER Errors Analysis

Table 4 shows the comparison of precision (**P**), recall (**R**) and $\mathbf{F_1}$ score among different teams, evaluated on the *Test-20* corpus. Here the exact match refers to the cases where a predicted entity is considered correct, only if the predicted type and boundary is exactly same as the gold entity.

Team	Word Representation	Features	Approach
Big Green	BERT	-	BERT fine-tuning
mgsohrab	PubMedBERT	-	PubMedBERT fine-tuning

Table 5: Summary of relation extraction systems designed by each team.

Whereas, in partial match, a predicted entity is considered correct if the predicted type is the same as the gold entity type and predicted entity boundary has some overlap with the gold entity boundary.

We observe that ensemble models with contextual word representations outperforms all other approaches by achieving 77.99 F_1 score in exact match (Team:BiTeM) and 81.75 F_1 score in partial match (Team:PublishInCovid19). Fine tuning of contextual word representation systems demonstrated quite competent performance with SciBERT-fine tuning being the best (Team:mgsohrab).

In Figure 2, we present an error analysis. Among the best performing models, the ensemble of transformer (Team:BiTeM) had significantly lower amount of 'over prediction' error (i.e., tagging a non-entity token as entity), compared to the system with ensemble of BiLSTM-CRFs (Team:PublishInCovid19).

3.2 RE Errors Analysis

Table 6 shows the comparison of precision (**P**), recall (**R**) and $\mathbf{F}_1$ score among the participant teams, evaluated on the *Test-20* corpus. Both of the teams utilized the gold entities and then predict the relations among these entities by fine-tuning contextual word representations. We observed that fine-tuning of domain related PubMedBERT, provides significantly higher performance compared to the general BERT fine-tuning. While examining the relation predictions from both of these systems, we found that system with PubMedBERT fine-tuning (Team:mgsohrab) resulted in significantly less amount of errors in every category (Figure 3).

	P	**R**	$\mathbf{F}_1$
mgsohrab	**80.86**	80.07	**80.46**
Big Green	45.42	**86.54**	59.57

Table 6: Results on extraction of 31 relation types from the *Test-20* dataset.

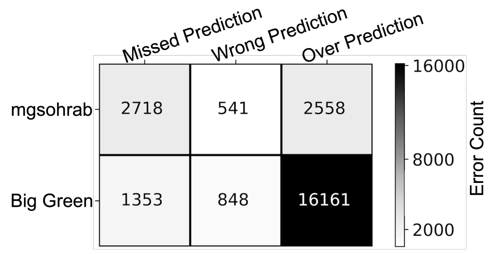

Figure 3: Summary of incorrectly predicted relations in each submitted systems.

4 Summary

In this paper, we presented a shared task for consisting of two sub-tasks: named entity recognition and relation extraction from the wet lab protocols. We described the task setup and datasets details, and also outlined the approach taken by the participating systems. The shared task included larger and improvised dataset compared to the prior literature (Kulkarni et al., 2018). This improvised dataset enables us to draw stronger conclusions about the true potential of different approaches. It also facilitates us analyzing the results of the participating systems, which in aids suggesting potential research directions for both future shared tasks and noisy text processing in user generated lab protocols.

Acknowledgement

We would like to thank Sydney Lee, Ethan Lee and Jaewook Lee for helping with data annotation.

References

Kaushik Acharya. 2020. KaushikAcharya at WNUT 2020 Shared Task-1: Conditional Random Field(CRF) based Named Entity Recognition(NER) for Wet Lab Protocols. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual String Embeddings for Sequence Labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*.

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

Maxwell Bates, Aaron J Berliner, Joe Lachoff, Paul R Jaschke, and Eli S Groban. 2017. Wet lab accelerator: a web-based application democratizing laboratory automation for synthetic biology. *ACS synthetic biology*.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. SciBERT: Pretrained Contextualized Embeddings for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *arXiv preprint arXiv:2007.15779*.

Saket Gupta, Aman Sinha, and Rohit Agarwal. 2020. DSC-IITISM at WNUT 2020 Shared Task-1: Name Entity Extraction from Wet Lab Protocol. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv preprint arXiv:2004.10964*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Sudeshna Jana. 2020. SudeshnaTCS at WNUT 2020 Shared Task-1: Name Entity Extraction from Wet Lab Protocol. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

Zhengbao Jiang, Wei Xu, Jun Araki, and Graham Neubig. 2020. Generalizing Natural Language Analysis through Span-relation Representations. In *Proceedings of the 2020 Conference of the Association for Computational Linguistics*.

Ayush Kaushal and Tejas Vaidhya. 2020. Winners at WNUT 2020 Shared Task-1: Domain specific BERT representation for Named Entity Recognition of lab protocol. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

Zhan Kecheng, Xiong Ying, Peng Hao, Yao, and LiQing Yao. 2020. BIO-BIO at WNUT 2020 Shared Task-1: Name Entity Extraction from Wet Lab Protocol.

Kabir Khan. 2020. kabir at WNUT 2020 Shared Task-1: Name Entity Extraction from Wet Lab Protocol. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

Julien Knafou, Nona Naderi, Jenny Copara, Douglas Teodoro, and Patrick Ruch. 2020. BiTeM at WNUT 2020 Shared Task-1: Named Entity Recognition over Wet Lab Protocols using an Ensemble of Contextual Language Models. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

Chaitanya Kulkarni, Wei Xu, Alan Ritter, and Raghu Machiraju. 2018. An Annotated Corpus for Machine Reading of Instructions in Wet Lab Protocols. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Fusataka Kuniyoshi, Kohei Makino, Jun Ozawa, and Makoto Miwa. 2020. Annotating and Extracting Synthesis Process of All-Solid-State Batteries from Scientific Literature. *arXiv preprint arXiv:2002.07339*.

Lukas Lange, Xiang Dai, Heike Adel, and Jannik Strötgen. 2020. B-NLP at WNUT 2020 Shared Task-1: Name Entity Extraction from Wet Lab Protocol. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*.

Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Chris Miller and Soroush Vosoughi. 2020. Big Green at WNUT 2020 Shared Task-1: Relation Extraction as Contextualized Sequence Classification. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

Nam Nguyen and Yunsong Guo. 2007. Comparisons of sequence labeling algorithms and extensions. In *Proceedings of the 24th international conference on Machine learning*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Mohammad Mahdi Abdollah Pour and Parsa Farinnia. 2020. mahab at WNUT 2020 Shared Task-1: Name Entity Extraction from Wet Lab Protocol. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

Utpal Kumar Sikdar, Bjorn Gamback, and M Krishana Kumar. 2020. IBS at WNUT 2020 Shared Task-1: Name Entity Extraction from Wet Lab Protocol. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

Janvijay Singh and Anshul Wadhawan. 2020. PublishInCovid19 at WNUT 2020 Shared Task-1: Entity Recognition in Wet Lab Protocols using Structured Learning Ensemble and Contextualised Embeddings. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

Mohammad Golam Sohrab, Khoa Duong, Makoto Miwa, and Hiroya Takamura. 2020. mgsohrab at WNUT 2020 Shared Task-1: Neural Exhaustive Approach for Entity and Relation Recognition Over Wet Lab Protocols. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

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*.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

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 biodesign automation*.

Alain C Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu H Nair, Philippe Schwaller, and Teodoro Laino. 2020. Automated extraction of chemical synthesis actions from experimental procedures.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*.

Qingcheng Zeng, Haoding Meng, Xiaoyang Fang, and Zhexin Liang. 2020. Fancy Man Launches Zippo at WNUT 2020 Shared Task-1: A Bert Case Model for Wet Lab Entity Extraction. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.