IITKGP at W-NUT 2020 Shared Task-1: Domain specific BERT representation for Named Entity Recognition of lab protocol

Tejas Vaidhya and **Ayush Kaushal**
Indian Institute of Technology, Kharagpur
iamtejasvaidhya@gmail.com, ayushk4@gmail.com

Abstract

Supervised models trained to predict properties from representations, have been achieving high accuracy on a variety of tasks. For instance, the BERT family seems to work exceptionally well on the downstream task from NER tagging to the range of other linguistic tasks. But the vocabulary used in the medical field contains a lot of different tokens used only in the medical industry such as the name of different diseases, devices, organisms, medicines, etc. that makes it difficult for traditional BERT model to create contextualized embedding. In this paper, we are going to illustrate the **System for Named Entity Tagging based on Bio-Bert**. Experimental results show that our model gives substantial improvements over the baseline and stood the **fourth runner up** in terms of F1 score, and **first runner up** in terms of Recall with just 2.21 *F1* score behind the best one.[1]

1 Introduction

A large amount of data is generated every year in the medical field. One of the most important generated data is the documentation of protocols. It provides individual sets of instructions that allow scientists to recreate experiments in their own laboratory. Most of them are written in Natural language which reduces its machine readability. The protocol gives a concise overview of the project which reduces its pre-processing needs but also make it less informative syntactically that eventually results in less accuracy.

Recent progress in Named Entity Recognition was made possible by the advancements of deep learning techniques used in natural language processing (NLP). For instance, Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Conditional Random Field (CRF)

(Namikoshi et al., 2017) have greatly improved performance in biomedical named entity recognition (NER) over the last few years. Bio-BERT(Lee et al., 2019) outperform all the other previous approaches with the help of BERT (Devlin et al., 2018) architecture pre-trained on Bio-medical texts (Giorgi and Bader, 2018; Habibi et al., 2017; Wang et al., 2018; Yoon et al., 2019).

In this paper, we are introducing our system for the NER tagging on the WLP dataset (Kulkarni et al., 2018). We use a variant of Bio-Bert (Lee et al., 2019). The primary motivation to use the model is it's medical vocabulary and features encoded in the pre-trained model.

2 Task Description and Data Set

Formally, the WNUT 2020 Shared Task-1 Named Entity Recognition, organized within, the 6th Workshop on Noisy User-generated Text (WNUT), 2020 (Tabassum et al., 2020) is a NER prediction task. It can be expressed as 'tokens-level' classification task mathematically as:

Let the sentence S be defined as:

$S = \{s_1, s_2, ..., s_n\}$

n is the number words in the sentences can be classified into the following label set

$y = \{l_1, l_2, l_3, ..., l_m\}$

where m is labels.

Given named entity of type XXX. Whenever two entities of type XXX are immediately next to each other, the first word of the second entity will be tagged B-XXX in order to show that it starts another entity and the entities inside B-XXX will be represented as I-XXX. For example, the sentence {nCoV-2019, sequencing, protocol} have the following labels {B-Reagent, B-Method, I-Method}.

DataSet: All of the protocols (Kulkarni et al., 2018) were collected from protocols.io using their public APIs by organising team. For the shared

---

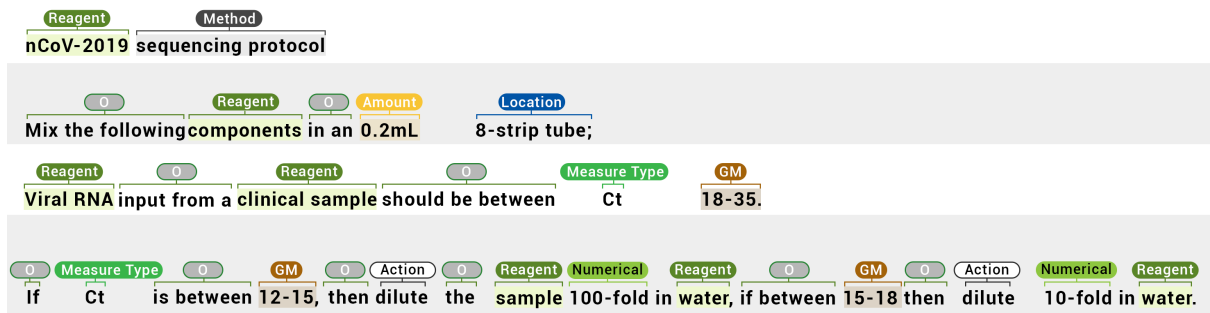[1] https://github.com/tejasvaidhyadev/
NER_Lab_Protocols

Figure 1: Visualisation of annotated dataset [2]

Task-1 W-NUT 2020: Named Entity Extraction, the annotation of 615 protocols are re-annotated using BART styled annotated protocols by 3 annotators with 0.75 inter-annotator agreement, measured by span-level Cohen's Kappa. The re-annotators incorporate missing entity-relations and also corrected the inconsistencies.

The task aims to create the system for Protocol-Named Entities Recognition (NER). The main difference that makes it difficult for traditional NER taggers is the vast vocabulary in medical filed and use of limited syntactic information. For instances "QIAprep Spin Miniprep" is device used in medical industry, but not present in our regular vocabulary that also makes it difficult for traditional NER tagger to learn.

Figure 1 provides the visualization of annotated datasets provided by WNUT 2020 Shared Task-1 .

3 Approach

In the section 3.1 we are going to define our proposed architecture and in section 3.2 briefly review the Bio-BERT (Lee et al., 2019) used for final submission and also different Domain specific BERT based model used for experiment as shown in the Table 1.

Baseline The organiser provided a simple Linear CRF model[3]. It utilized simple gazetteers and hand-crafted feature to predict the entities from the test data. We replaced it with our proposed BERT based Architecture as describe in the section below.

3.1 Architecture

As described in figure 2, we first sub-word tokenize each token of sentences, using BERT's word-piece tokenizer of Huggingface library and pass it

through different domain specific BERT models or BERT Transformer stacks (scibert, biobert, bert-based, bert-large etc) to extract contextualised representation (Beltagy et al., 2019; Lee et al., 2019; Devlin et al., 2018). We then select the representation of first sub-word token for each word and use simple Linear or Dense layer with the softmax activation function as classifier to get probability on the labels from the contextualised representation.

3.2 Bio-BERT

Bio-BERT (Lee et al., 2019) is a contextualized language representation model, based on BERT, a pre-trained model that is trained on different combinations of general & biomedical domain corpora. According to Lee et al. (2019), just like its parental model BERT, it is also capable of capturing contextualized bidirectional representations. Thus it has outperformed existing architectures in most of the Named Entity Recognition tasks within the biomedical domain by using a limited amount of dataset. We hypothesize that such domain-specific bidirectional representations are also critical for our task. Bio-BERT are pre-trained on the following different datasets {*Wiki + Books, Wiki + Books + PubMed, Wiki + Books + PMC, Wiki + Books + PubMed + PMC*}.

we again hypothesize to achieved best performance in *PubMed*(comprises more than 30 million citations for biomedical literature from MEDLINE, life science journals, and online books.) trained dataset because of its linguistic similarity with protocols. For instance both of them contains medical procedure to reproduce medical experiment.

Other Models used in Experiment We also used other domain specific BERT for experimentation using the same architecture as discussed in section 3.1 with replacement of BERT Models in place of BioBERT (as shown in figure 2)

---

[2]Image Source: Github repository of WNUT 2020 Shared Task-1 NER.

[3]https://github.com/jeniyat/WNUT_2020_NER/tree/master/code/baseline_CRF

Models	F1-score	Recall	Precision
`biobert_v1.1_pubmed`	79.10	79.72	78.61
`biobert_v1.0_pubmed_pmc`	79.02	79.51	79.02
`scibert_uncased`	77.66	79.60	76.00
`bert-large-cased`	77.79	78.74	77.10
`bert-large-uncased`	75.50	77.39	73.79
`bert-base-cased`	78.05	79.29	76.87
`Baseline`	74.39	73.32	75.49

Table 1: Shows the results of test set provided by shared task organisers during experimental and details of the experimental setting is describe in section 4
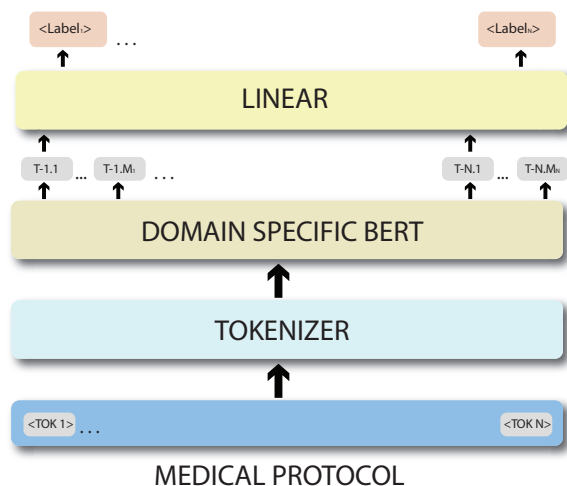

Figure 2: The TOK X represents the token of sentence where X ∈ N and N is the length of sentence. T-I.J represents the Ith tokens and J^{th} subtokens and we used WordPiece tokenizer for all of our models.

- Sci-BERT (Beltagy et al., 2019) : A pretrained contextualized embedding model based on BERT to address the lack of high-quality, large-scale labeled scientific data.

- BERT$_{\{large/base,cased/uncased\}}$:BERT (Devlin et al., 2018) It is designed to train deep bidirectional representations by jointly conditioning on both left and right context in all layers. Language models have demonstrated that rich, unsupervised pre-training is an integral part of many language understanding systems. Hence, we try fine-tuning BERT to obtain better results on this task.

4 Comparison and Discussion

We experimented with different BERT models as shown in Table 1. We avoid any preprocessing other than the Bert specific tokenization, as it may result in loss of crucial semantic information in text.

We also assumed protocols are relatively less noisy compare to other crowd source data with compact sentences.

4.1 Experimental setting

We keep maximum length of input sentence to 512 to consider Long sentences in protocols. For all the base models (12-layer, 768-hidden, 12-heads, 110M parameters) including our Bio-BERT model, we train all for 8 epoch with batch size 16. Large models(24-layer, 1024-hidden, 16-heads, 340M parameters.) are trained for 4 epochs with batch size 16. We early stop the models using the valid set. The dropout probability was set to 0.1 for all layers. Optimization is done using Adam(Kingma and Ba, 2014) with a learning rate of 1e-5. The remaining hyperparameters were kept same as Devlin et al. (2018). We used the PyTorch (Paszke et al., 2019) implementation of BERT from Huggingfaces tranformers (Wolf et al., 2019) library.

For selecting best models in experimental phase (i.e. before release of test set) we use split of 60/20/20 for train, dev and test respectively. Our final submission, used a 70/30 split for train/valid set of initial data and Bio-BERT(Lee et al., 2019) model and split sentence with more than 512 tokens to two more sentence to get desired model's input sentence length. To evaluate the performance of the system, an evaluation script along with the dataset was provided by the organizers[4].

4.2 Results and Inferences

Our Bio-Bert(Lee et al., 2019) based model performed best of all the models because of domain specific knowledge. Our model performed extremely well on final test set as shown in Table 2 and stood 4th *runner* up in term of F1 score and

---

[4] `https://github.com/jeniyat/WNUT_2020_NER/tree/master/code/eval`

Models	F1-score	Recall	Precision
`biobert_v1.1_pubmed (partial match)`	79.54	77.43	81.76
`biobert_v1.0_pubmed_pmc (complete match)`	74.91	72.93	77

Table 2: Results on the held-out test set provided by shared task organisers on final submission

Tokens	correct labels	predicted labels
standard	B-Reagent	B-Reagent
T4	I-Reagent	B-Device
DNA	I-Reagent	I-Reagent
Ligase	I-Reagent	I-Reagent

Table 3: Error arises due to consideration of token level classification

Bio-Med Terms	subword-tokenized
acetyltransferase	['ace','ty','lt','ran','s','fer','ase']
Hematoxylin	['He','mat','ox','yl','in']
sulfanilamide	['su','lf','ani','lam','ide']
ddH2O	['d','d','H','2','O']
lBiotin-16-UTP	['l','B','iot','in','-','16','-','U','TP']

Table 4: Illustration of inefficient sub-tokenization of Bio-Med words

1st runner up in term of recall out of 13 teams participated in the competition.

Inferences: Use of uncased large BERT results in significant loss of about 2.29 F1 score in comparison of cased-large BERT (Devlin et al., 2018), clearly shows importance of syntactic nature of protocols. We observed better performance, after increase in both training and validation set for our final submissions indicating inability of model to fully capture the representation due to fine tuning on limited data.

5 Error Analysis

1. BERT tokenizer is not efficient on Bio-Medical text as illustrated in Table 4 . Its vocabulary does not consists of Bio-medical words and it is not trained on domain specific setting. Hence makes it difficult for BERT model to learn the encoding based on poorly sub tokenized word. The possible solution will be training tokenizers on both biomedical and general text sets.

2. Treatment of task as token level classification problem results in incorrect detection of intermediate Entity as illustrated in Table 3. Hence evidenced in decrease of F1 score in complete match compare to partial match in our submission provided by organisers.

3. Use of Nomenclature, Scientific formula, abbreviations makes it difficult for Pre-trained language models to generalised with limited fine tuning data. Though with the help of contextual details it is observed that the BERT was able to correctly predict scientific formula

at some places. For instance, in the sentence {Add, 5gm, SDS}, *SDS* was correctly labelled as *Reagent* by our model.

6 Conclusion and Future Work

In this paper, we present our system for the Named Entity recognition for Bio-medical protocols a for Shared Task at W-NUT Task-1 2020. We built upon the recent success of Pre-trained language models and apply them for protocols. Our System achieves close to state-of-art performance on this task.

As future work, we will try to experiment with XLNet(Yang et al., 2019) and different ensembling between the models and would like to extend the work of Clark et al. (2019) by performing layer by layer Analysis of BERT.

Acknowledgments

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

---

[5]https://www.linkedin.com/in/santosh-t-y-s-s-39227b116/

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does bert look at? an analysis of bert's attention. In *Black-BoxNLP@ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805Comment: 13 pages.

John M Giorgi and Gary D Bader. 2018. Transfer learning for biomedical named entity recognition with neural networks. *Bioinformatics*, 34(23):4087–4094.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

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 (NAACL)*.

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*, 36(4):1234–1240.

D. Namikoshi, M. Ohta, A. Takasu, and J. Adachi. 2017. Crf-based bibliography extraction from reference strings using a small amount of training data. In *2017 Twelfth International Conference on Digital Information Management (ICDIM)*, pages 59–64.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Jeniya Tabassum, Wei Xu, and Alan Ritter. 2020. WNUT-2020 Task 1: Extracting Entities and Relations from Wet Lab Protocols. In *Proceedings of EMNLP 2020 Workshop on Noisy User-generated Text (WNUT)*.

Xuan Wang, Yu Zhang, Xiang Ren, Yuhao Zhang, Marinka Zitnik, Jingbo Shang, Curtis Langlotz, and Jiawei Han. 2018. Cross-type biomedical named entity recognition with deep multi-task learning. *Bioinformatics*, 35(10):1745–1752.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.

Wonjin Yoon, Chan So, Jinhyuk Lee, and Jaewoo Kang. 2019. Collabonet: collaboration of deep neural networks for biomedical named entity recognition. *BMC Bioinformatics*, 20:249.