IITKGP AT W-NUT 2020 SHARED TASK-1
DOMAIN SPECIFIC BERT REPRESENTATION
FOR NAMED ENTITY RECOGNITION OF LAB PROTOCOL

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

Task Introduction

Lab protocols specify steps in performing a lab procedure. They are noisy, dense, and domain-specific. Automatic or semi-automatic conversion of protocols into machine-readable format benefits biological research.

In this task, 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. The main difference that makes it difficult for traditional NER taggers is the vast vocabulary in medical filed and use of limited syntactic information.

Dataset

- We used the data provided by the share task organisers, containing the Named Entity for Lab Protocols.
- All of the protocols (Kulkarni et al.,2018) were collected from protocols.io using their public APIs by organising team.
- For the Task, the annotations are re-annotated using BART styled annotated protocols by 3 annotators with 0.75 inter-annotator agreement, measured by span-level Cohen's Kappa.

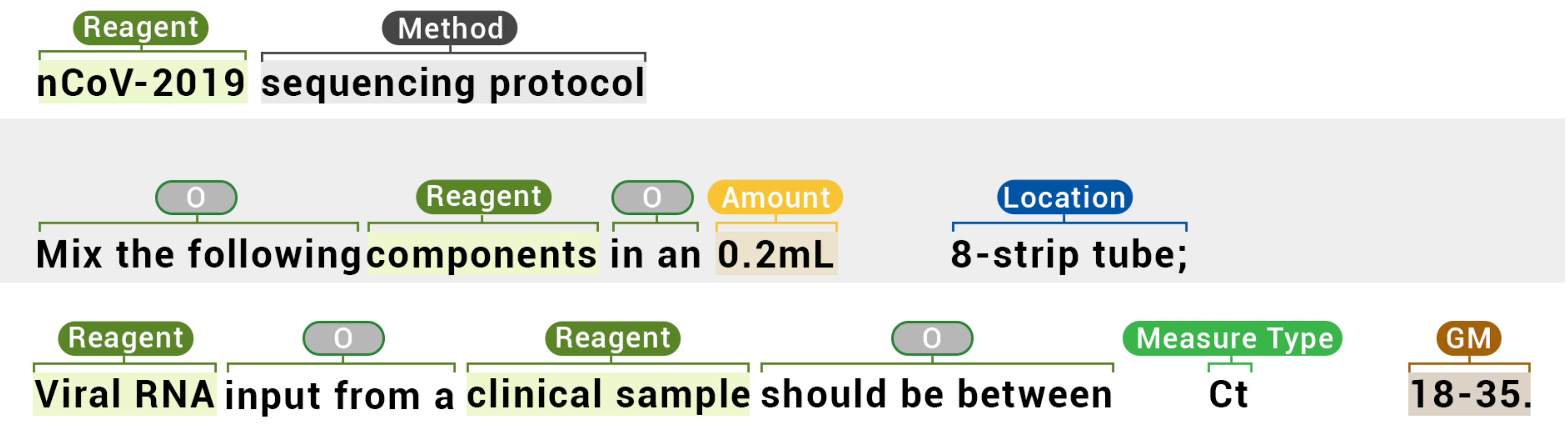
Reagent
Method
nCoV-2019 sequencing protocol
O
Reagent
O
Amount
Location
Mix the following components in an 0.2mL 8-strip tube;
Reagent
O
Reagent
O
Measure Type
GM
Viral RNA input from a clinical sample should be between Ct 18-35.

Figure 1: Visualisation of annotated dataset

Approach

- We used different version and domain-specific BERT to learn the best pre-trained representation.
- **Baseline**:
  - The organiser provided a simple Linear Conditional Random Fields model.
  - It utilized simple gazetteers and handcrafted feature to predict the entities from the test data.
- We replaced it with our proposed BERT based Architecture.

Architecure

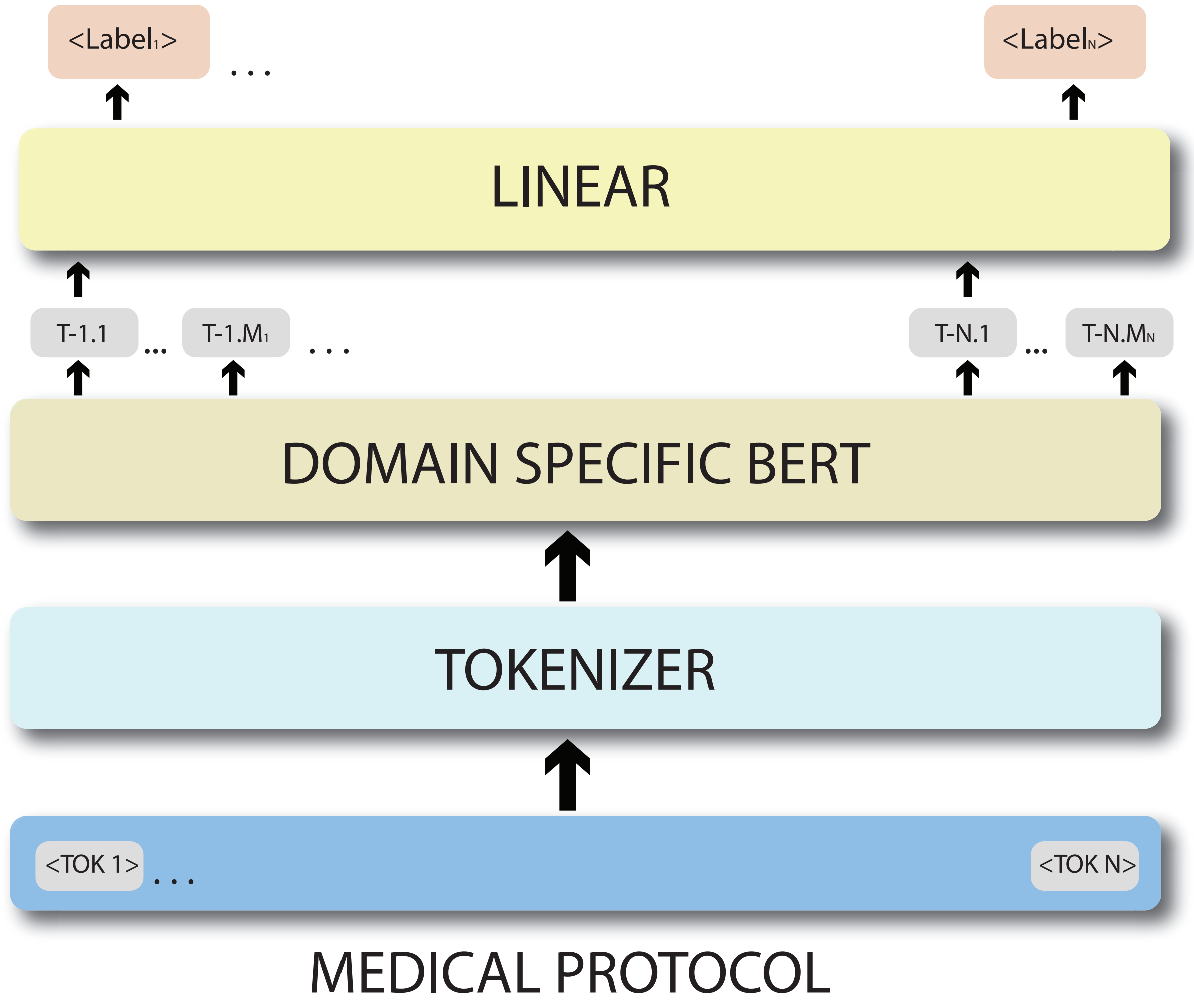

Figure 2: The $TOKX$ represents the token of sentence where $X \in N$ and N is the length of sentence.

we used the following pre-trained models for our experiments:

- **Bio-Bert**: Bio-BERT are pre-trained on the following different datasets Wiki + Books, Wiki + Books + PubMed, Wiki + Books + PMC, Wiki + Books + PubMed + PMC.
- **Sci-BERT**: A pretrained contextualized embedding model based on BERT to address the lack of highquality, large-scale labeled scientific data.
- **BERT** :It is designed to train deep bidirectional representations by jointly conditioning on both left and right context in all layers.

Results

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

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

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 and stood **4th runner up** in term of **F1 score** and **1st runner up** in term of **recall** out of 13 teams participated in the competition.

Error Analysis

- BERT tokenizer is not efficient on Bio- Medical text as illustrated in Table 3. Its vocabulary does not consists of Bio-medical words.

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 3: Illustration of inefficient sub-tokenization of Bio-Med words

- Treatment of task as token level classification problem results in incorrect detection of intermediate Entity as illustrated in Table 4.

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 4: Error arises due to consideration of token level classification

- 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.