



**Course : Advance NLP**

**Final Report Submission**

# Named Entity Recognition with Small Strongly Labeled and Large Weakly Labeled Data

**Submitted to :**

**Dr. Manish Kumar Srivastava**

**Team :**

Vaibhav Singh Tomar (2020101058)
Mashrukh Islam (20161137)

## 1. Introduction

Strongly labeled supervised data achieves quality performance, however, owing to practical considerations, models are trained using weakly labeled data which results in underperforming systems. The referred work explores this problem due to weak supervision in the field of NER by proposing NEEDLE a three-stage training procedure in the face of small supervised and large weakly labeled data. The first stage involves usual continual pretraining on the domain-specific unlabeled data. The crux of the NEEDLE framework lies in the second stage which introduces task-specific pretraining with the aid of weak label completion and noise-aware loss function. In the third stage, fine tuning on a strongly labeled dataset achieves good results on the task.

## 2. Dataset And Exploratory Analysis

We evaluated the proposed framework on two different domains: E-commerce query domain and Biomedical domain.

**For Biomedical** NER, we use three popular benchmark datasets: BC5CDR-Chem, BC5CDR Disease (Wei et al., 2015), and NCBI-Disease (Doğan et al., 2014). These datasets only contain a single entity type. We use the pre-processed data in BIO format from Crichton et al. (2017) following BioBERT (Lee et al., 2020) and PubMedBERT (Gu et al., 2020)

| Biomedical Domain | | | | | | |
|---|---|---|---|---|---|---|
| BC5CDR Chem | 5K | 5K | 5K | 11M | 92.08 | 77.40 |
| BC5CDR Disease | 5K | 5K | 5K | 15M | 94.46 | 81.34 |
| NCBI Disease | 5K | 1K | 1K | | | |

Table 1: Data Statistics

Link to the datasets have been attached below:

https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/

https://huggingface.co/datasets/ncbi_disease

https://paperswithcode.com/dataset/bc5cdr

## 3. Architecture

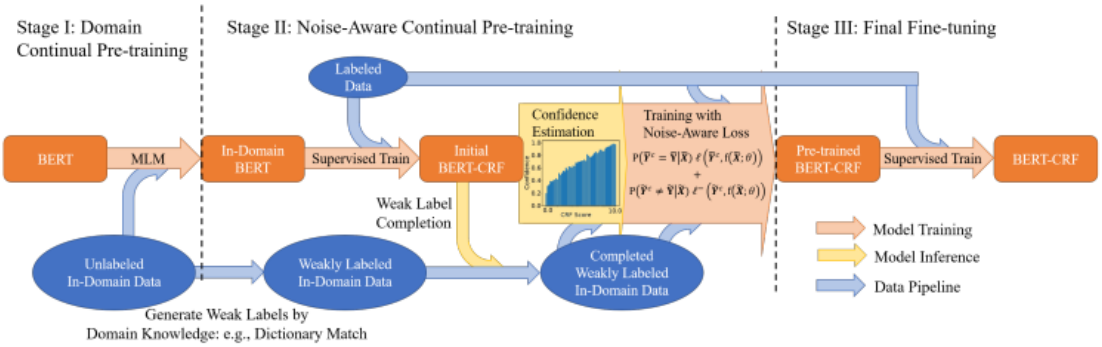


Figure 1: Three-stage NEEDLE Framework.

## 4. Project Implementation

I first started with looking for datasets, listed in the paper, So there are 2 datasets which were used namely E-commerce one and the other Biomedical one.

The E-commerce dataset is not made public by amazon, only mentioned in the paper, where as the other dataset which is mentioned in the papers which of PubMed from National Government health website which keep updating every year, So the exact specific dataset on which the paper was published was not available, but the updated dataset was available which we took.

The Original Dataset is about 20.6 GB which was difficult to preprocess, which we later on shrink it because of computational unavailability.

From hereon we downloaded the dataset and first pre-processed it, cleaned it, and refined it. It was pure text corpus, which we referred as unlabelled data.

We then create weak labels for them using the chemical and disease dictionary mentioned. We then conduct domain continual masked language model pre-training on the large in-domain unlabelled data. Till this point, Data Preprocessing has been done and We have started implementing the model architecture. Stage I is almost done.

We downloaded a BERT Model for baseline comparisons, The hugging face link to that model is provided below:

https://huggingface.co/dmis-lab/biobert-v1.1

After that we started Stage II, In the second stage, we use the knowledge bases to convert the unlabeled data to weakly labeled data to generate weak labels for the unlabeled data. And after that We fine tuned the model on the strongly labelled dataset.

## 5. BaseLine Comparison:

We contrasted the performance gains due to the NEEDLE approach against a generally followed BERT+CRF model as the baseline.

## Performance Benchmark

### BioMedical NER

| Method (F1) | BC5CDR-chem | BC5CDR-disease | NCBI-disease |
|---|---|---|---|
| BERT | 89.99 | 79.92 | 85.87 |
| bioBERT | 92.85 | 84.70 | 89.13 |
| PubMedBERT | 93.33 | 85.62 | 87.82 |
| **Ours** | **91.38** | **82.57** | **88.53** |

## 7. Link To Model and Dataset

The whole 20.6 GB dataset, has been pre-processed and stored in ada-GPU 50 scratch folder. Since it was not possible to pre-process that big or even a small subset of that dataset in normal system of ada storage.

Link to directory:

Gpu 50: ./scratch/flugeltomar/testing/unlabelled_data

The dataset used for model training and model link is provided below:

https://iiitaphyd-my.sharepoint.com/:f:/g/personal/vaibhav_tomar_students_iiit_ac_in/EntWZLrPmD1Lk13O6CPnL6wBrWMaYVb83BAJkJCCV4BGhw?e=rkmEMP

## 8. References:


- https://github.com/vstflugel/Advance-NLP-Named-Entity-Recognition-with-Small-Strongly-Labeled-and-Large-Weakly-Labeled-Data
- https://huggingface.co/datasets/ncbi_disease
- https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/
- https://paperswithcode.com/dataset/bc5cdr
- https://github.com/amzn/amazon-weak-ner-needle/tree/main
- https://www.youtube.com/watch?v=2XUhKpH0p4M
- https://www.youtube.com/watch?v=uKPBkendlxw
- https://aclanthology.org/2021.acl-long.140/