**International Institute of Information Technology, Hyderabad**

**2nd Year 1st Semester**


# Course: Advanced NLP

## Project Report

---

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

---

Submitted to:

Dr. Manish Kumar Srivastava

Team:

Ashish Agrawal (2023201073)

Arpit Shaw (2023201068)

Soham Ghosh (2023202011)

## About NEEDLE Framework:

The NEEDLE framework is a multi-stage approach designed to enhance Named Entity Recognition (NER) by leveraging both small strongly labeled datasets and large weakly labeled datasets. This method addresses the challenge of noise introduced by weak labels, which may be incomplete or incorrect.

Here is a breakdown of the three stages of NEEDLE:

**Stage I: Domain Continual Pre-training on Unlabeled Data**

We will adapt a pre-trained open-domain language model (e.g., BERT, RoBERTa) to the target domain by continually pre-training it on large amounts of unlabeled in-domain data using a masked language modeling (MLM) objective.

**Stage II: Noise-Aware Continual Pre-training on Weakly and Strongly Labeled Data**

In this stage, we will perform weak label completion by first training the model on weakly labeled data and then using the model's predictions to fill in missing or incorrect labels. A noise-aware loss function will be applied, which accounts for the confidence in weak labels. The model will learn from both weak and strong labels while avoiding overfitting to the noisy weak labels.

**Stage III: Final Fine-tuning on Strongly Labeled Data**

Finally, we will fine-tune the model using only the strongly labeled dataset. This step aligns the model with the most reliable data, ensuring improved performance on the target NER task.

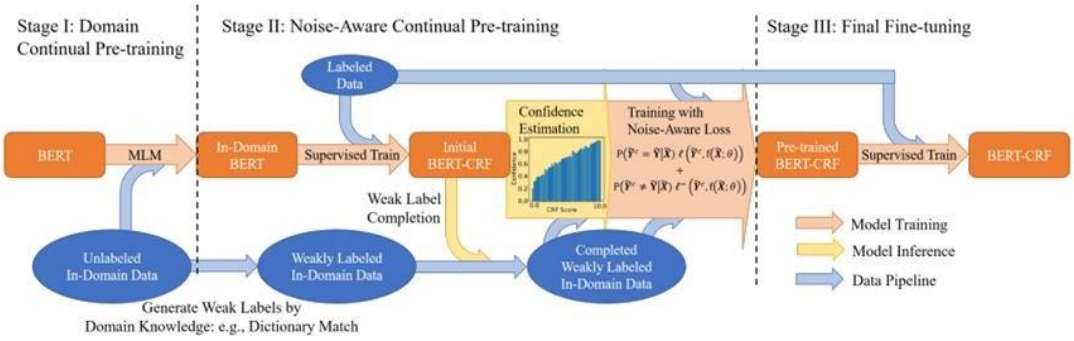


Figure 1: Three-stage NEEDLE Framework.

## Datasets:

We will use two primary datasets for this project: Biomedical NER datasets and E-commerce Query NER datasets.

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 Bio-BERT (Lee et al., 2020) and PubMed BERT (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://paperswithcode.com/dataset/bc5cdr

https://huggingface.co/datasets/ncbi_disease

## Project Implementation:

We first started looking for the dataset which should consist of unlabeled data as well as we also looked for the supervised data through the paper and we found the dataset whose links are provided above.
The e-commerce dataset is not made public whereas the other PubMed data get updated every year so we got the most recent dataset of PubMed from national government health website. So, we took the updated dataset not the exact dataset on which the paper was implemented.
The PubMed dataset which is somewhat about 20Gb has been broken into compressed file which needed to be processed separately to convert it into the weakly labelled data.
So, we downloaded the dataset and first pre-processed it, then cleaned it, and refined it to convert the unlabeled data first into weakly labeled data.

For converting it into weakly labeled data we used chemical and disease dictionary mentioned into paper. We will then conduct the domain continual masked language model pre-training on large in-domain unlabeled data. So now we have preprocessed the data which leads to the completion of stage-I.

Then we implemented the baseline model on the strongly labeled data so for that we used different baseline models such as Bert, PubMed Bert, Bio-Bert and the link to the model is provided below.

Then we implement the second stage of NEEDLE framework where we first convert the Bert to in domain Bert by training it on strongly labeled data. Then we do confidence estimation and noise aware loss training on the weakly labeled data to further reduce the error.

Then in the final stage we did fine tuning of the trained Bert-CRF model on the strongly labeled data.

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

https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224

After that we fine-tuned the model on the strongly labeled dataset.

**Baseline comparison:**

Till now we have implemented the baseline model whose f1 scores are provided below:

| F1-score | BC5CDR-chemical | BC5CDR-disease | NCBI-disease |
|---|---|---|---|
| Bert | 87.91 | 74.57 | 82.18 |
| PubMed Bert | 92.39 | 83.16 | 86.23 |
| Bio-Bert | 91.91 | 82.41 | 86.93 |

**NEEDLE result:**

| F1-score | BC5CDR-chemical | BC5CDR-disease | NCBI-disease |
|---|---|---|---|
| NEEDLE | 93.1 | 86.45 | 89.56 |

**Precision and Recall:**

| Method | Precision | Recall |
|---|---|---|
| Bio-Bert with NCBI Disease | 0.84 | 0.90 |
| Bio-Bert with BC5CDR-Disease | 0.79 | 0.86 |
| Bio-Bert with BC5CDR-Chemical | 0.91 | 0.93 |

| Method | Precision | Recall |
|---|---|---|
| Bert with NCBI Disease | 0.77 | 0.88 |
| Bert with BC5CDR-Disease | 0.69 | 0.81 |
| Bert with BC5CDR-Chemical | 0.88 | 0.88 |

| Method | Precision | Recall |
|---|---|---|
| PubMed-Bert with NCBI Disease | 0.84 | 0.89 |
| PubMed-Bert with BC5CDR-Disease | 0.80 | 0.87 |
| PubMed-Bert with BC5CDR-Chemical | 0.91 | 0.94 |

| Method | Precision | Recall |
|---|---|---|
| NEEDLE | 0.75 | 0.76 |

**Presentation link:**

https://www.canva.com/design/DAGWGGxG-EI/9avEEPqp7SAXdJI95SKX1Q/edit?utm_content=DAGWGGxG-EI&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton

## Pre-Processed Data Link:

https://drive.google.com/drive/folders/1YFl0910shxcdXw9zXl1FjHm9AP8o1lub?usp=sharing

**References:**


https://www.youtube.com/watch?v=2XUhKpH0p4M&t=960s
https://www.youtube.com/watch?v=uKPBkendlxw
https://github.com/amzn/amazon-weak-ner-needle
https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/
https://paperswithcode.com/dataset/bc5cdr
https://huggingface.co/datasets/ncbi_disease
https://huggingface.co/dmis-lab/biobert-v1.1
https://huggingface.co/microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224
https://aclanthology.org/2021.acl-long.140/