Detecting Biomedical Named Entities in COVID-19 Texts

Shaina Raza [1 2] **Brian Schwartz** [1 2]

Abstract

The application of the state-of-the-art biomedical named entity recognition task faces a few challenges: first, these methods are trained on a fewer number of clinical entities (e.g., disease, symptom, proteins, genes); second, these methods require a large amount of data for pre-training and prediction, making it difficult to implement them in real-time scenarios; third, these methods do not consider the non-clinical entities such as social determinants of health (age, gender, employment, race) which are also related to patients' health. We propose a Machine Learning (ML) pipeline that improves on previous efforts in three ways: first, it recognizes many clinical entity types (diseases, symptoms, drugs, diagnosis, etc.), second, this pipeline is easily configurable, reusable and can scale up for training and inference; third, it considers non-clinical factors related to patient's health. At a high level, this pipeline consists of stages: pre-processing, tokenization, mapping embedding lookup and named entity recognition task. We also present a new dataset that we prepare by curating the COVID-19 case reports. The proposed approach outperforms baseline methods on four benchmark datasets with macro-and micro-average F1 scores around 90, as well as using our dataset with a macro-and micro-average F1 score of 95.25 and 93.18 respectively.

1. Introduction

In recent years, the number of biomedical documents (research papers, case reports, electronic health records, and clinical notes) has increased dramatically. MEDLINE, a comprehensive database of medical articles, contains approximately 28 million articles to date (MEDLINE, 2021). Due to COVID-19 research, hundreds of articles have been published in the past two years (Wang & Lo, 2021). To keep up with the increasing demand for biomedical knowledge, large-scale data management is necessary. In its current state, it is very challenging for researchers to manage and infer information from unstructured (free) texts (Campos et al., 2012). These challenges include parsing the scientific text, extracting key information, categorizing the articles, and facilitating efficient content discovery. Text mining (Tan et al., 1999) is a subtask of Natural Language Processing (NLP) that converts free texts into a format suitable for data analysis and to build machine learning (ML) models.

In the modern healthcare industry, there is also a substantial increase in Electronic Health Record (EHR) data (Toscano et al., 2018). It may include patient conditions, medications, demographic data, medical history, and laboratory reports. Before these EHRs can be utilized for research purposes, they must be de-identified.

The task of identifying and categorizing key information (entities such as a person, an organization, or an event) in the text is known as Named Entity Recognition (NER) (Nadeau & Sekine, 2007), and it is a key technique in text mining. The NER task can be used in the biomedical domain to identify biomedicine entities such as genes, diseases, species, chemicals, and so on (Cho & Lee, 2019). The results of named entities can be used for a variety of downstream tasks, including question answering system, drug-drug interaction analysis, gene identification and information extraction. The state-of-the-art work (Efroni et al., 2020) in biomedical NER focuses on a small number of named entities (disease, genes, proteins, etc.). However, there are many entities to consider, such as disease, diagnosis, medical concepts, risks, vital signs, and so, that need to be identified from texts, which is a motivation for this research.

According to the Healthy People 2030 initiative (of Health), non-clinical factors such as social determinants of health (SDoH) (McNeely et al., 2020), live, work, and grow that influence the health of populations. At a broader level, SDoHs refer to the distribution of wealth, power, and resources that have long-term effects on individual health outcomes and lead to health disparities (McNeely et al., 2020). For instance, the effect of food insecurity on the patients' health and the effect of substandard housing on mental health. In this study, we also focus on SDoHs, which is a relatively

[1]Public Health Ontario (PHO), Toronto, ON, Canada [2]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada. Correspondence to: Shaina Raza <shaina.raza@oahpp.ca>.

Proceedings of the 1st Workshop on Healthcare AI and COVID-19, ICML 2022.

under-researched area in healthcare and AI.

The purpose of this research is to study the clinical and non-clinical named entities from biomedical texts. We refer to both clinical and non-clinical entities as "biomedical entities" throughout this paper. We propose a trainable ML pipeline that includes many phases, such as pre-processing, tokenization, embedding lookup, a deep neural network for the NER task; a NER converter to convert identified named entities into user-friendly representations; and a de-identifier to de-identify patients' personal information based on identified named entities (name, location, date, etc.). We summarize our contributions as:

- We propose and develop a Biomedical NER Pipeline (BNP), to identify biomedical entities from the scientific texts. This pipeline consolidates and explains ML best practices with a variety of features that can be used as-is or as a starting point for further customization and enhancement.
- We create a new dataset by curating a large number of COVID-19 case reports, and scientifically parsing the text from these case reports. A case report describes a patient's symptoms, diagnosis, treatment, and follow-up. We also annotate a portion of this dataset with biomedical entities to create a gold-standard dataset that is used to train and evaluate the named entity model.
- After identifying the named entities (name, address, etc.), we de-identify the patients' personal information in the case report to comply with the Personal Health Information Protection Act (PHIPA) (Nosowsky & Giordano, 2006). In this work, we do not have access to real patients' personal information, so we use the fake identifiers to build and test this module.
- We set up this pipeline using the Spark NLP configurations (Kocaman & Talby, 2021) that allows us to scale up in clusters while maintaining distributed data processing principles. Spark NLP supports in-memory distributed data processing for both training and inference processes in real-time.

We compare the effectiveness of our NER approach to state-of-the-art methods on publicly available benchmark datasets and our COVID-19 case reports dataset.

2. Related Work

Traditional NER methods only consider specific entities (e.g., persons, organizations, locations, etc.) (Nadeau & Sekine, 2007). Biomedical NER (Campos et al., 2012) is the task of identifying entities in the biomedical domain, such as chemical compounds, genes, proteins, viruses, disorders, drugs, adverse effects, diseases, DNAs and RNAs. In the state-of-the-art of biomedical NER, most of the research (Efroni et al., 2020; Goyal et al., 2018) focuses on general approaches to named entities that are not specific to the biomedical field. On the other hand, there are some works (Eltyeb & Salim, 2014; Lee et al., 2020) that focus solely on biomedical and chemical NER, however, they do not cover many clinical entities, such as diseases, symptoms, clinical procedures and such. SDoHs also a major impact on people's health, and well-being and are related to health outcomes, which is rather an under-explored research area in bio-medicine research.

In the last few years, there has been a dramatic increase in biomedical data (1). Recently, because of the COVID-19 surge, there is much increase in biomedical data that is difficult to read, even more so when the urgency of time and the number of patients is increasing exponentially. Due to the critical nature of comprehending and fully utilizing this biomedical data, several NLP tasks are initiated. These tasks include Biomedical Question Answering (Raza et al., 2022), CORD-19 challenge (Wang et al., 2020), and TREC-COVID challenge (Roberts et al., 2021). To perform these tasks, it is, therefore, necessary to accommodate a prior process of biomedical NER task.

According to a 2016 survey, about 95% of U.S. hospitals use EHRs (Toscano et al., 2018). Case reports (Rison et al., 2013) also contain patients' data that can be used as a substitute for EHRs (Hummel & Evans, 2016) and are distributed for free for research purposes. The term "de-identification" refers to the process of removing or replacing personal identifiers in such a way that re-establishing a link between this information should not be possible Some studies employ de-identification as a sub-task of the biomedical NER task (Fabregat et al., 2019), where patient personal entities are recognized first and are then de-identified.

Usually, the NER tasks are considered as sequence-labeling problems, where words in a given phrase are tokens that can be given appropriate labels. Consideration of correlations represented by the best joint probability between neighbouring labels and the full sequence of labels is useful for sequence-labeling. The Conditional Random Field (CRF) models (Lafferty et al., 2001) are usually useful for sequence-labeling where we can jointly decode label sequences using a CRF layer.

CRF models (Lafferty et al., 2001), and Structured Support Vector (SVM) (Tsochantaridis et al., 2005) are commonly used models for NER, biomedical NER and de-identification tasks. Deep learning models based on recurrent neural networks (RNN) and Convolutional Neural Network (CNN) are also used for biomedical named entities and de-identification purpose (Yang et al., 2019). In recent times, BioBERT (Lee

et al., 2020), SciBERT (Beltagy et al., 2019) and related Transformer-based models are also used to identify named entities from biomedical texts.

In this work, we also use deep learning-based methods to build a pipeline for the biomedical NER and de-identification tasks. We extend the standard biomedical NER to identify many named entities.

3. Approach

We develop a ML pipeline, **B**iomedical **N**amed entity recognition **P**ipeline (BNP), shown in Figure 1 . This pipeline takes raw data, pre-processes it, and applies algorithms to recognize and classify biomedical entities into predefined classes. Next, we discuss each stage of this pipeline in detail.

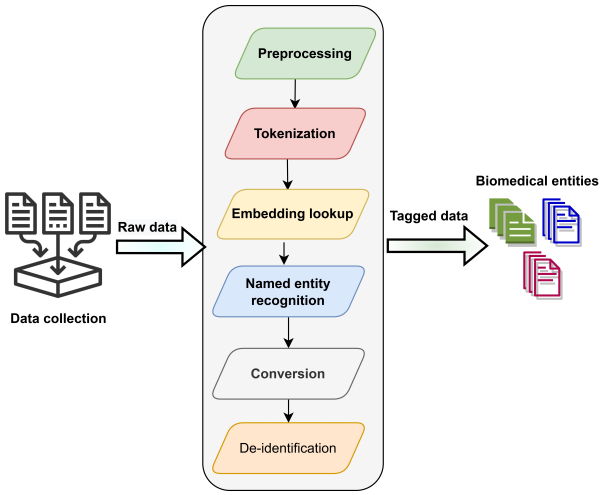

Figure 1. Proposed pipeline for biomedical named entity recognition task

3.1. Data Collection

The input to the BNP is the data that is collected from different sources. We feed some benchmark biomedical datasets, and our dataset on COVID-19 case reports to the pipeline.

Benchmark datasets: We use the following benchmark datasets: JNLPBA (Kim et al., 2004), NCBI-Disease (Doğan et al., 2014), BC5CDR (Li et al., 2016), BC2GM (Smith et al., 2008).

These datasets are readily available in CoNLL-2003 format here . CoNLL has become a prototypical standard for building algorithms that recognize named entities in the texts (Sang & De Meulder, 2003). We performed additional processing to convert the datasets into IOB (Inside-Outside-Before) scheme (Zhai et al., 2017). The IOB format is a tagging format in computational linguistics (e.g., NER) and is used for sequence labelling. It is a facility provided by our pipeline.

Our COVID-19 case reports dataset: We have collected the clinical case reports from different journals (Lancet, BMJ, AMJ, Clinical Medicine and other related journals) that are standardized according to the CAseREports (CARE) guidelines (Rison et al., 2013). The inclusion criteria are given below:

- We include only the PubMed Central (PMC) case reports.
- We specify English as the language for the case reports.
- We specify the timeline between March 20, 2021 and March 20, 2022 for data collection.
- We exclude many early-pandemic case reports, as the disease symptoms, diagnosis, drugs, and vaccination information were unclear at that time. After scraping the PDFs of these case reports, we use Apache Tikka toolkit to extract metadata (authors' names, DOI, journal name, case report title) and full texts from PDF documents. After completing these steps, we found around 4500 case reports.

The dataset details are as: *JNLPBA (JNL)* with gene/ protein entity types consists of 35,336 annotations from 2,404 abstracts. *NCBI* dataset with disease entity type consisting of 6,881 annotations from 793 abstracts. *BC5CDR (BC5)* with chemical entity type consisting of 15,935 annotations from 1,500 scientific articles. *BC2GM (BC2)* with gene/ protein entity type have 24, 583 annotations from 20,000 sentences. Our case reports (*Cases*) data with many clinical and non-clinical entities have around 25,000 annotations from 500 case reports. The actual size of the case reports dataset is 4500 but we use 500 case reports for annotations, which according to research heuristics (Kocaman & Talby, 2021; Chen et al., 2020) is a good size to start training the model.

Gold-standard dataset: Gold-standard dataset (Ogren et al., 2008) means a corpus of text or a set of documents that are manually annotated with the labels. We use the John Snow Labs annotation lab [1] to annotate around 400 case reports and prepare it in the CONLL (Sang & De Meulder, 2003) format to construct a gold-standard dataset; which according to research (Snow et al., 2008), is good number to begin training an NLP model. We train the BiSLTM-CNN-CRF algorithm inside the pipeline with this gold-standard dataset to initiate the biomedical NER task.

[1] https://www.johnsnowlabs.com/annotation-lab/

3.2. Proposed pipeline

Our proposed BNP, shown in Figure 1 consists of various stages that are discussed next.

Pre-processing: The first stage in BNP is the data pre-processing stage that is handled by a node 'pre-processor'. The input to this stage is the data from data collection stage. The pre-processor pre-processes the input data and detects the sentence boundaries in each document. Then, it transforms the data into a format that is readable by the next stage in the pipeline. The output from this stage is the set of pre-processed documents.

Tokenization: The tokenized data from the previous stage (i.e., tokenization) goes into the embedding lookup stage, which is handled by the embedding lookup node. We have used the BERT-based clinical embeddings pre-trained on PubMed corpora and MEDLINE. This embedding lookup node maps tokens to vectors, it can also download other pre-trained embeddings (such as Glove, BERT, BioBERT, etc.,). The output from this stage is word embeddings corresponding to each word in the document.

Embedding lookup: The tokenized data from the previous stage (i.e., tokenization) goes into the embedding lookup stage, which is handled by the embedding lookup node. We have used the BERT-based clinical embeddings[2] pre-trained on PubMed corpora and MEDLINE. This embedding lookup node maps tokens to vectors, it can also download other pre-trained embeddings (such as Glove, BERT, BioBERT, etc.,). The output from this stage is word embeddings corresponding to each word in the document.

Named entity recognition: This stage identifies biomedical entities in the documents. It is an algorithm that is based on Bi-directional Long short-term memory (BiLSTM) - Convolutional Neural Network (CNN) - Conditional Random Field (CRF) (Huang et al., 2015) model. We refer to this model as BiLSTM-CNN-CRF model. We modify the vanilla BiLSTM-CNN-CRF for the data loader so that it can work with any input by converting it into CoNLL and then into IOB encoding scheme. We show the working of BiLSTM-CNN-CRF in Figure 2 and explain its work next.

As shown in Figure 2, the BiLSTM-CNN-CRF algorithm takes the sequence of words as $S = [w_1, w_2, \ldots, w_N]$ as input, where w_i refers to the one-hot representation of the *ith* word in the sequence. This input goes to the first layer, which is the embedding layer. The embedding layer converts a sentence from a sequence of characters into a sequence of dense vectors. An embedding matrix $\mathrm{E} \in R^{D \times V}$ is used to map each character into a dense vector, where *D* is the embedding dimension and *V* is the vocabulary size.

The output of embedding layer is a sequence of vectors

[2]https://github.com/ncbi-nlp/bluebert

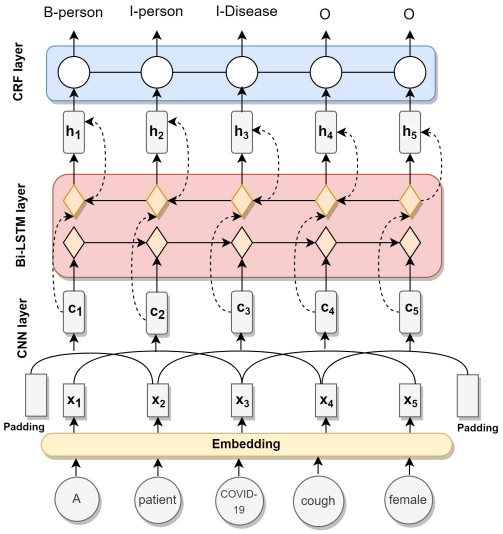

Figure 2. Proposed pipeline for biomedical named entity recognition task

$S = [x_1, x_2, \ldots, x_N]$, where $\mathrm{x_i} \in R^{D \times V}$ and x_i refers to the dense vector representation of word w_i. We use the pre-trained embeddings that are loaded and provided by the embedding lookup node in this layer.

The second layer in this NER node is a CNN network that is used to capture local information within given words in a biomedical context. Each position in the sequence has sliding windows, and CNN performs a transformation for each sliding window. The contextual representation c_i of the i_{th} character is learned by using the CNN filter, as shown in Equation 1:

$$c_i = f\left(w^T \bigoplus x_{[i \pm \frac{K-1}{2}]}\right) \tag{1}$$

where $x_{[i \pm \frac{K-1}{2}]})$ represents the concatenation of embeddings of characters. We use the Rectified Linear activation Unit (ReLU) as an activation function *f*. The contextual representation c_i is the concatenation of the outputs of all filters at this position. The output of CNN layer is $C = [c_1, c_2, \ldots, c_N]$, where $c_i \in R^M$, *M* refers to the number of filters in CNN layer.

The third layer in the model is the Bi-LSTM network that is used to learn hidden representations of characters for tokens in a sequence using all previous contexts (in both directions). The hidden representation h_i is a concatenation of contexts in both directions. The output of Bi-LSTM layer is $h = [h_1, h_2, \ldots, h_N]$, where *S* refers to the dimension of hidden states in LSTM.

The fourth layer on the top of the Bi-LSTM network is the CRF layer (Ma & Hovy, 2016) The input to the CRF layer is $h = [h_1, h_2, \ldots, h_N]$ generated by the Bi-LSTM layer,

where h refers to the sequence of hidden states. CRF is a conditional probability distribution model mostly used in sequence labelling tasks to generate new tags based on previously labelled tags (Ma & Hovy, 2016). In any NER task, the neighboring labels strong depend on the target label. For example, I-Disease (I for inside) label usually follows B-Disease (B for before), but it cannot follow B-SYMPTOM or I-SYMPTOM. Thus, it is useful to jointly decode the labels of characters in a sequence rather than decode them independently.

The output of the CRF layer is $y = [y_1, y_2, \ldots, y_N]$, where y refers to the sequence of labels. In this work, the biomedical entities are the labels. A *tanh* layer on top of the BiLSTM layer is added to predict the confidence scores (CS) for every word with each of the possible labels as the output score of the network, as shown in Equation 2:

$$CS_i = \tanh\left(\mathrm{W_c} h'_i + b_\mathrm{c}\right) \tag{2}$$

where W_c and b_c refers to model parameters. In training, we use the negative log likelihood function over all training samples to calculate the loss function $\mathcal{L}$, which is shown in Equation 3 as:

$$\mathcal{L}_{\text{BNP}} = -\sum_{\mathrm{s}\in\mathrm{S}} \log\left(p\left(y_\mathrm{s} \mid h_\mathrm{s}; \theta\right)\right) \tag{3}$$

where S refers to the set of sentences in training data, p denotes the probability and θ refers to the parameters during training.

Conversion: This stage converts the IOB representation of named entities to a user-friendly representation, by associating the tokens of recognized entities and their labels. This stage is handled by NER convertor node. Each output from this stage is a 'chunk' that is a tagged portion of sentence into named entities.

De-identification In this stage, we employ the data obfuscation technique, which is a process that obscures the meaning of data (Bakken et al., 2004). For example, to replace identified names with different fake names or to mask some data value $\langle 04-04-2022 \rangle$ with $\langle DATE \rangle$. This component provides HIPAA or PHIPA compliance when dealing with text documents containing any protected health information. We use the pre-trained de-identification model (JSL) from John Snow Labs inside the pipeline to de-identify patients' records.

3.3. Biomedical named entities

We get biomedical named entities as the output of the BNP. We include a number of clinical and non-clinical entities that we finalized after reviewing the relevant literature (Caufield et al., 2019; Johnson et al., 2016). These entities are:

Clinical entities Admission (patient admission status), oncology (tumor/cancer), blood pressure, respiration (shortness of breath), dosage (medicine), vital signs, symptoms, kidney disease, temperature (body), diabetes, vaccine, time of symptom (days, weeks), obesity, pregnancy, BMI, height (of patient), heart disease, pulse, hypertension, drug name, drug ingredient, hyperlipidemia, cerebrovascular disease, disease syndrome disorder, treatment, clinical department, weight (of patient), admission/ discharge (from hospital), modifier (modifies current state), external body part, test, strength, route, test result, drug.

Non-Clinical entities Name (of patient), location, date, relative date, duration, relationship status, social status, family history (family members, alone/ with family/ homeless), employment status, race/ethnicity, gender, sexual orientation, diet (food type, nutrients, minerals), alcohol, smoking

4. Experiments and Results

4.1. Experimental settings

We use PyTorch for the implementation of models. In addition, we use the Spark NLP pipeline to construct the BNP (pipeline), which allows us to scale up in clusters and supports in-memory distributed data processing for faster training and inference. We run our experiments on Google Colab Pro (NVIDIA P100, 24 GB RAM, 2 x vCPU) and used Apache Spark NLP in local mode (no cluster) to integrate the components of the ML pipeline. We specify the following hyper-parameters as shown in Table 1. We use Grid search to get the optimal values for the hyper-parameters and early stopping to overcome possible over-fitting.

We also tried different pre-trained embeddings in the embedding lookup component, such as glove100d, word2vec and BERT embeddings and find better performance with BERT embeddings pre-trained on PubMed and MEDLINE corpora. We have divided all datasets into training, validation, and test sets, with a 70:15:15 ratio. We used the Stratified 5-Folds cross-validation (CV) strategy for train/test split if original datasets do not have an official train/test split.

Evaluation metrics Following the standard practice (Chen et al., 2015) to evaluate NER tasks, we use the following metrics:

- Micro-average F1 (mi) measures F1-score of aggregated contributions of all classes.
- Macro-average F1 (ma) adds all the measures (Precision, Recall, or F-Measure) and divides with the number of labels, which is more like an average

Baseline methods We test the performance of our BNP approach against the following methods:

SciBERT (Beltagy et al., 2019): we use the implementation of allenai/scibert-base pre-trained on biomedical data with

Table 1. Hyperparameters used.

HYPERPARAMETER	OPTIMAL VALUE (VALUES USED)
LEARNING RATE	1.E-03 (1.E-02,..., 3.E-04)
BATCH SIZE	64 (8, 16, 32, 64, 128)
EPOCHS	30 ({2, 3, . . . , 30})
LSTM STATE SIZE	200 (200, 250)
DROPOUT RATE	0.5 ({0.3, 0.35, ..., 0.7})
OPTIMIZER	ADAM
CNN FILTERS	2 (2,3,4,5)
HIDDEN SIZE	768
EMBEDDING SIZE	128
MAX SEQ LENGTH	512
WARMUP STEPS	3000

785k vocabulary.

BioBert (Lee et al., 2020): BioBERT is a pre-trained language model for biomedical text mining. We use the BioBERT-base-cased with following versions:

BioBert v1.0 pre-trained on 200k PubMed articles.

BioBERT v1.1 pre-trained on 1M PubMed articles.

BioBERT v1.2 pre-trained on 1M PubMed articles in the same way as BioBERT v1.1 but includes a language modelling (LM) head.

CT-BERT (Müller et al., 2020): it is a BERT-large-uncased model, pre-trained on Twitter messages on the topic of COVID-19.

BiLSTM-CRF (Akbik et al., 2018): we use a standard BiLSTM-CRF architecture that relies on contextual string embeddings.

All the baselines are trained on the datasets mentioned in Table 1. Each baseline is tuned to its optimal hyper-parameter setting and the best results for each baseline are reported.

4.2. Results

Comparison with baseline methods We report the results of all methods on all datasets using macro (ma) average F1 in Table 2 and micro (mi) -average F1 scores in Table 3. These scores show the percentage values. Bold means highest and italic means second highest performance.

Overall, these results in Table 2 and 3 show that our BNP approach achieves the best performance on four public biomedical benchmarks (NCBI Disease, BC5CDR, JNLPDP, BC2GM) as well as on our case-reports designed specifically for biomedical named entities. This demonstrates the generalizability of our methodology across different types of datasets.

This outstanding performance of our approach is attributed

Table 2. Test results using micro F1 average

	NCBI	BC5	JNL	BC2	CASES
CT-BERT	62.67	62.91	60.27	62.82	68.16
SCIBERT	81.15	80.72	77.13	76.78	76.23
BILSTM-CRF	83.32	83.92	79.23	78.04	81.23
BIOBERT 1.0	86.01	84.56	78.68	85.28	85.87
BIOBERT 1.1	88.52	87.15	79.39	86.16	86.27
BIOBERT 1.2	*89.12*	*87.81*	*83.34*	*86.45*	*86.88*
BNP (OURS)	**91.12**	**89.12**	**90.13**	**89.15**	**93.14**

Table 3. Test results using macro F1 average

	NCBI	BC5	JNL	BC2	CASES
CT-BERT	63.14	63.24	61.15	63.23	68.72
SCIBERT	82.13	79.88	80.65	80.13	78.29
BILSTM-CRF	84.12	84.02	83.56	79.32	78.10
BIOBERT 1.0	79.10	78.90	79.00	78.13	72.18
BIOBERT 1.1	85.89	87.10	*87.18*	*85.45*	*87.78*
BIOBERT 1.2	*86.78*	*87.89*	86.07	85.15	86.98
BNP (OURS)	**91.14**	**89.14**	**89.01**	**90.23**	**95.25**

to two important things: (1) the embedding lookup component that can load the domain-specific pre-trained language model (we use clinical BERT embeddings) to get the relevant embeddings. (2) Our approach stacks together various ML components or nodes (Figure 1) as a directed acyclic graph (sequence of execution steps), where each node prior to NER node contributes to identifying the biomedical entities. We see the biggest performance boost when our pipeline is tested on our case reports dataset that is annotated with many biomedical named entities.

Our approach achieves the best micro F1 score of 93.14 on our dataset (around 52 entities), 91.12 on NCBI Disease (disease entity), 89.12 on BC5CDR (chemicals), 89.15 on BC2GM (gene/proteins) and 90.13 on JNLPDP (gene/proteins) dataset. We see similar patterns of highest performance by our pipeline for macro F1 scores with a performance of 95.25 using our dataset.

The BioBERT model also shows a competitive performance (after our model) in these results. We find that BioBERT achieves better performance on disease entities (NCBI), followed by chemical (BC5CDR) and then gene/proteins entities (BC2GM and JNLPDP). It performed very well on our case reports dataset, probably because it has rich clinical embeddings. Among the variants of BioBERT, we see the overall better performance of BioBERT v1.2 than its other predecessors, except for a few places, where BioBERT v1.1 marginally outperforms BioBERT v1.2. The better performance of BioBERT v1.2 attributes to its training method, which is the same way as BioBERT v1.1 but includes a

language model head, which can be useful for probing. A probing task (Perone et al., 2018) is a classification problem that focuses on the simple linguistic properties of embeddings. In this work, we use the clinical BERT embedding pre-trained on PubMed and Medline data, which shows at least 1-3% better results than BioBERT pre-trained on PubMed abstracts.

We also observe the performance of BiLSTM-CRF model in identifying many diseases, chemical and gene/proteins entities in these experiments. The BiLSTM-CRF, though not as deeper as a BERT model, performs better than SciBERT and CT-BERT. This is probably because SciBERT is initially trained on scientific data (not clinical), its performance is somehow compromised on biomedical entities. In the same way, CT-BERT is pre-trained on social media data, so the meanings of entities are different from pure biomedical entity types.

Our approach, BioBERT v1.1 and v1.2 perform better than simple BiLSTM-CRF baseline. Our modified BiLSTM-CNN-CRF algorithm performs better than BiLSTM-CRF baseline, probably because, we are using pre-trained biomedical embeddings (BiLSTM-CRF uses Glove embeddings). We are using a deeper neural network - with more layers than standard BiLSTM-CRF. The BioBERT is also pre-trained on huge amounts of biomedical data and is a self-sufficient model to determine biomedical NER, so it outperforms BiLSTM-CRF baseline. SciBERT and CT-BERT relatively lower performance could be attributed to the fact that it does not cover as much training and inference on biomedical entities as our approach and BioBERT.

Although we fine-tune each baseline method to its optimal hyper-parameter settings, we anticipate that the relatively low scores of these baselines on our case reports dataset can be attributed to the following: (i) absence of a training dataset for training new biomedical, and (ii) different training/test set splits used in previous works that were unavailable.

In comparison to previous research, our method can recognize a wide variety of clinical and non-clinical entity types. We extract many entries related to medical risk factors (hypertension, kidney, diabetes, etc.), patients' personal information (age, gender, geography), SDoH (diets, race, income) and other clinical entity types such as underlying disease, tissue, and organ systems. Most biomedical-focused projects concentrate on chemicals, proteins, and genes; however, our pipeline is quite adaptable and can identify many identities. This is demonstrated by the fact that when we train our pipeline on other datasets, it performed best. When we trained our pipeline on the case reports dataset, it again performed the best.

Effectiveness of BNP on case reports We give a random snippet from a COVID-19 case report to our pipeline and show the confidence scores for the predicted biomedical entities. The BiLSTM network in the BNP predicts the confidence scores (Equation **??**) for every word with possible labels. The expectation over here is that for a given confidence score, the model should predict with a higher confidence score. The results are shown in Table 4.

Table 4. Confidence scores of predicted biomedical entities (sen. for sentence, beg. for beginning conf. for confidence score).

SEN.	BEG.	END	CHUNK	ENTITY	SCORE
0	2	12	85 YEAR OLD	AGE	1.00
0	14	18	WOMAN	GENDER	0.98
0	32	43	ICU	CLINICAL DEPT	0.93
0	109	134	FEVER	SYMPTOM	0.91
0	156	160	COUGH	SYMPTOM	1.00
0	233	244	PRIOR 5 DAYS	RELATIVE DATE	0.93
1	247	249	SHE	GENDER	0.99
1	261	264	MILD	MODIFIER	0.78
1	266	273	DIARRHEA	SYMPTOM	0.82

Table 5. Most used clinical entities in 100 case reports

DRUG	DISEASE	SYMPTOM
ANTIBIOTICS	CORONAVIRUS	COUGH
DOBUTAMINE	CARDIOGENIC	COLD
OSELTAMIVIR	COVID-19	CONGESTION
FLUIDS	PNEUMONIA	ABDOMINAL
SALINE	HEPATITIS	HEMORRHAGE

As seen in Table 4, our pipeline can predict many biomedical entities from the input text. For brevity reasons, we only present a snippet of a case report, so these predicted entities do not encompass all the biomedical entities that we have defined.

We also show five most common clinical entities predicted using our approach from 100 case reports, due to space limitation, we only show a few top entities in Table 5.

We demonstrate a few non-clinical entities with their results in Table 6 and find some factors that can be analyzed to study the social impacts on population health.

We show the de-identification of personal information in Figure 3.

Finally, we show a sample prediction of our BNP pipeline on a case report [3] in Figure 4.

[3] https://casereports.bmj.com/content/13/5/e235861

Table 6. Most used non-clinical entities in 100 case reports

RACE	RELATIONSHIP	AGE	GENDER
CAUCASIAN	SINGLE	18-YEAR	FEMALE
ASIAN	MARRIED	59-YEAR-	HE
BLACK	DIVORCED	TEENAGER	HIS
WHITE	PARTNER	OLDER	MAN
LATIN	WINDOWED	YOUNGER	SHE

index	sentence	deidentified
0	A 73-year-old woman came to the Fever Clinic of the First Hospital.	A 73-year-old woman came to the Fever Clinic of the GENESYS REGIONAL MEDICAL CENTER - HEALTH PARK.
1	The name of patient is Oliveira who lives in Cocke County Baptist Hospital .	The name of patient is Rayna Buttery who lives in CHRISTUS MOTHER FRANCES HOSPITAL - SULPHUR SPRINGS .
2	0295 Keats Street.	P.O. Box 101.
3	Phone 111-347-837.	Phone 06435 14 64 15.

Figure 3. De-identification task

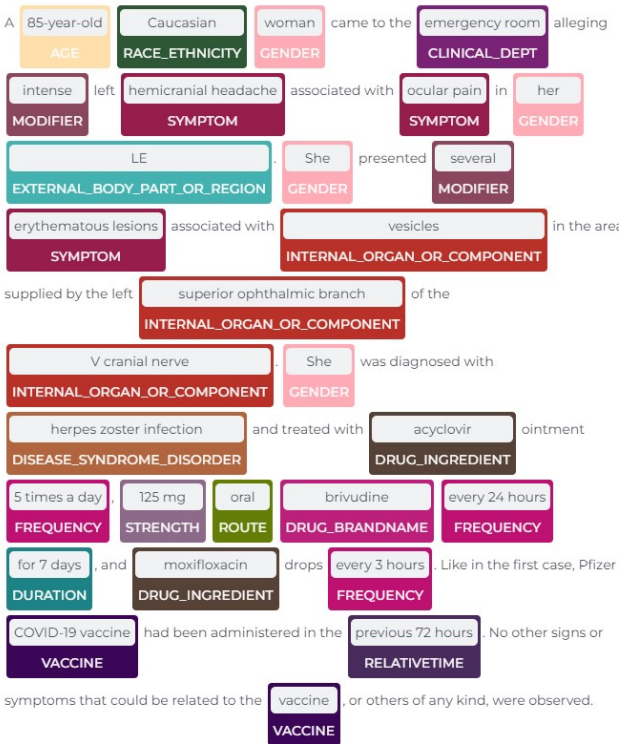

Figure 4. Predictions on a case report

5. Conclusion

In conclusion, this paper presents a ML pipeline for biomedical NER task that consists of a number of nodes stacked together. We use BiLSTM-CNN-CRF model plus BERT-based embeddings to detect biomedical entities. The results show that using contextualized word embedding pretrained on biomedical corpora significantly improves the results. We evaluated the performance of our approach on five datasets (four benchmark datasets and one own developed case reports dataset) and our approach achieves the best results compared to the baselines.

Limitations and future work: This work is based on the curation of case reports and the biases related to study eligibility criteria, identification and selection of studies can be a limitation. So far, we chose only English as the language, which may have omitted many useful literatures from the corpus. One direction, in this regard is to have rigorous research methods to determines the quality of literature. For now, we don't have access to any EHR, so we cannot determine the validity of the de-identification component, we have uses fake identifiers to simulate the de-identification process, which suffice the purpose but does not represent the real-time patients' data. Some points that we plan to consider to implement in the future are:

We plan to add additional layers that may emphasize key patterns and words that are decisive for the identification of the named entities We plan to use a Transformer architecture (Devlin et al., 2018) that consists of a positional encoding layer that can compute the linear distance between words. This is significant because the model will contain additional information about the entities, their dependencies and cues that are important for the entity types. We strongly encourage the inclusion of medical professionals in the annotation process. This is one of the most important findings that we gathered from this research. In this way, the model will have a higher degree of confidence in the inaccurate predictions it generates. We also plan to do error analysis on inaccurate predictions generated by the model for the biomedical named entities

We also recommend annotating with standard biomedical terminologies and mappings, (such as those from Unified Medical Language System (UMLS), Medical Subject Headings (MeSH), International Classification of Diseases (ICD): ICD-9, ICD-10, Systematized Nomenclature of Medicine (SNOMED) terms) rather than developing a custom terminology from scratch. Well-maintained biomedical terminologies are often the result of ongoing expert effort, which is our long-term goal. We also plan to incorporate interoperability in various contexts and use cases for the biomedical research.

Acknowledgements

This work was supported by the Canadian Institutes of Health Research (CIHR)-Institute of Population and Public Health (IPPH) and Public Health Ontario (PHO).

References

Akbik, A., Blythe, D., and Vollgraf, R. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics*, pp. 1638–1649, 2018.

Bakken, D. E., Rarameswaran, R., Blough, D. M., Franz, A. A., and Palmer, T. J. Data obfuscation: Anonymity and desensitization of usable data sets. *IEEE Security & Privacy*, 2(6):34–41, 2004.

Beltagy, I., Lo, K., and Cohan, A. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

Campos, D., Matos, S., and Oliveira, J. L. Biomedical named entity recognition: a survey of machine-learning tools. *Theory and Applications for Advanced Text Mining*, 11:175–195, 2012.

Caufield, J. H., Zhou, Y., Bai, Y., Liem, D. A., Garlid, A. O., Chang, K.-W., Sun, Y., Ping, P., and Wang, W. A comprehensive typing system for information extraction from clinical narratives. *medRxiv*, pp. 19009118, 2019.

Chen, Q., Leaman, R., Allot, A., Luo, L., Wei, C.-H., Yan, S., and Lu, Z. Artificial intelligence (ai) in action: Addressing the covid-19 pandemic with natural language processing (nlp). *arXiv preprint arXiv:2010.16413*, 2020.

Chen, Y., Lasko, T. A., Mei, Q., Denny, J. C., and Xu, H. A study of active learning methods for named entity recognition in clinical text. *Journal of biomedical informatics*, 58:11–18, 2015.

Cho, H. and Lee, H. Biomedical named entity recognition using deep neural networks with contextual information. *BMC bioinformatics*, 20(1):1–11, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Doğan, R. I., Leaman, R., and Lu, Z. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, 2014.

Efroni, S., Song, M., Labatut, V., Emmert-Streib, F., Perera, N., and Dehmer, M. Named Entity Recognition and Relation Detection for Biomedical Information Extraction. *Frontiers in Cell and Developmental Biology — www.frontiersin.org*, 1:673, 2020. doi: 10.3389/fcell.2020.00673. URL www.frontiersin.org.

Eltyeb, S. and Salim, N. Chemical named entities recognition: a review on approaches and applications. *Journal of cheminformatics*, 6(1):1–12, 2014.

Fabregat, H., Duque, A., Martínez-Romo, J., and Araujo, L. De-identification through named entity recognition for medical document anonymization. In *IberLEF@ SEPLN*, pp. 663–670, 2019.

Goyal, A., Gupta, V., and Kumar, M. Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21–43, 2018.

Huang, Z., Xu, W., and Yu, K. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

Hummel, J. and Evans, P. Producing accurate clinical quality reports for population health: A delivery system-oriented approach to report validation. 2016.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

JSL. nlu/deidentification_model_overview.ipynb at master · johnsnowlabs/nlu · github. URL https://nlp.johnsnowlabs.com/2021/01/29/deidentify_enriched_clinical_en.html.

Kim, J.-D., Ohta, T., Tsuruoka, Y., Tateisi, Y., and Collier, N. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pp. 70–75. Citeseer, 2004.

Kocaman, V. and Talby, D. Spark nlp: natural language understanding at scale. *Software Impacts*, 8:100058, 2021.

Lafferty, J., McCallum, A., and Pereira, F. C. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wiegers, T. C., and Lu, Z. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016, 2016.

Ma, X. and Hovy, E. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*, 2016.

McNeely, C. L., Schintler, L. A., and Stabile, B. Social determinants and covid-19 disparities: Differential pandemic effects and dynamics. *World Medical & Health Policy*, 12(3):206–217, 2020.

MEDLINE. MEDLINE Overview, 2021. URL https://www.nlm.nih.gov/medline/medline_overview.html.

Müller, M., Salathé, M., and Kummervold, P. E. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*, 2020.

Nadeau, D. and Sekine, S. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30 (1):3–26, 2007.

Nosowsky, R. and Giordano, T. J. The health insurance portability and accountability act of 1996 (hipaa) privacy rule: implications for clinical research. *Annu. Rev. Med.*, 57:575–590, 2006.

of Health, U. D. Healthy people 2030 — health.gov. URL https://health.gov/healthypeople.

Ogren, P., Savova, G., and Chute, C. Constructing evaluation corpora for automated clinical named entity recognition. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

Perone, C. S., Silveira, R., and Paula, T. S. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*, 2018.

Raza, S., Schwartz, B., and Rosella, L. C. CoQUAD: a COVID-19 question answering dataset system, facilitating research, benchmarking, and practice. *BMC Bioinformatics*, 23(1):210, 2022. ISSN 1471-2105. doi: 10.1186/s12859-022-04751-6. URL https://doi.org/10.1186/s12859-022-04751-6.

Rison, R. A., Kidd, M. R., and Koch, C. A. The CARE (CAse REport) guidelines and the standardization of case reports, 2013.

Roberts, K., Alam, T., Bedrick, S., Demner-Fushman, D., Lo, K., Soboroff, I., Voorhees, E., Wang, L. L., and Hersh, W. R. Searching for scientific evidence in a pandemic: An overview of trec-covid. *Journal of Biomedical Informatics*, 121:103865, 2021.

Sang, E. F. and De Meulder, F. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

Smith, L., Tanabe, L. K., Kuo, C.-J., Chung, I., Hsu, C.-N., Lin, Y.-S., Klinger, R., Friedrich, C. M., Ganchev, K., Torii, M., et al. Overview of biocreative ii gene mention recognition. *Genome biology*, 9(2):1–19, 2008.

Snow, R., O'connor, B., Jurafsky, D., and Ng, A. Y. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 254–263, 2008.

Tan, A.-H. et al. Text mining: The state of the art and the challenges. In *Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases*, volume 8, pp. 65–70, 1999.

Toscano, F., O'Donnell, E., Unruh, M. A., Golinelli, D., Carullo, G., Messina, G., and Casalino, L. P. Electronic health records implementation: can the European Union learn from the United States? *European Journal of Public Health*, 28(suppl_4):cky213—-401, 2018.

Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y., and Singer, Y. Large margin methods for structured and interdependent output variables. *Journal of machine learning research*, 6(9), 2005.

Wang, L. L. and Lo, K. Text mining approaches for dealing with the rapidly expanding literature on covid-19. *Briefings in Bioinformatics*, 22(2):781–799, 2021.

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R., Liu, Z., Merrill, W., et al. Cord-19: The covid-19 open research dataset. *ArXiv*, 2020.

Yang, X., Lyu, T., Li, Q., Lee, C.-Y., Bian, J., Hogan, W. R., and Wu, Y. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. *BMC medical informatics and decision making*, 19(5): 1–9, 2019.

Zhai, F., Potdar, S., Xiang, B., and Zhou, B. Neural models for sequence chunking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.