A Project Report for Natural Language Processing

**Doctor Shrewd**

Submitted to Manipal University, Jaipur

Towards the partial fulfilment for the Award of the Degree of

**BACHELOR OF TECHNOLOGY**

In Data Science

2022-2023

By

Pranjal Paira 209309016



Under the Guidance of

Dr. Vivek Kumar Verma

**Department of Information Technology**

**School of Information Technology**

**Manipal University Jaipur**

**Jaipur, Rajasthan**

(Nov-2022)

**CERTIFICATE**

Date: 05/11/2022

This is to certify that the project work entitled **“Doctor Shrewd”** submitted by **Pranjal Paira (209309016)** in fulfilment for the requirements of the award of Bachelor of Technology Degree in Data Science at Manipal University Jaipur is an authentic work carried out by them under my supervision and guidance. To the best of my knowledge, the matter embodied in the project has not been submitted to any other University / Institute for the award of any Degree.

Dr. Vivek Kumar Verma Dr. Pankaj Vyas

Associate Professor

& Project Guide HOD

Department of IT Department of IT

Manipal University Jaipur Manipal University Jaipur

**ABSTRACT**

**Named Entity Recognition (NER)** is a vital step in medical information extraction, especially Electronic Health Records (EHRs). Proper extraction of medical entities such as disease and medications can automate the process of EHR coding as well as considerably improve the filtering of EHR resulting in **better extraction of medical information**.

In this context we have such a data set in which a lot of text is written related to the medical domain. There are a lot of diseases that can be mentioned in the entire dataset and their related treatments are also mentioned implicitly in the text.

We can build an algorithm to map the diseases and their respective treatment. In this project, a NER tagger is proposed, covering mainly multiple entity types, medications and diseases. The model used is Conditional Random Fields (CRF) for sequence tagger. The train dataset is used to train the Continuous Random Field (CRF) model, and the test dataset is used to evaluate the built model.

**CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.No.** |  | **Page** |
| i) | Certificate | i |
| ii) | Abstract | ii |
| 1. | Introduction | 1-3 |
| 1.1 | Motivation | 2 |
| 1.2 | Objectives | 3 |
| 2. | Literature Review | 4 |
| 3. | Design and Implementation | 5-9 |
| 4. | Results | 10 |
| 5. | Conclusion | 11 |
|  | References | 12 |

1. **INTRODUCTION**

**Named Entity Recognition (NER)** is a vital step in medical information extraction, especially Electronic Health Records (EHRs). Proper extraction of medical entities such as disease and medications can automate the process of EHR coding as well as considerably improve the filtering of EHR resulting in **better extraction of medical information**.

For example,

“The patient was a 62-year-old man with squamous cell **lung cancer**,

which was first successfully treated by a combination of radiation

therapy and **chemotherapy**."

As we can see in this text, a person with a non-medical background cannot understand the various medical terms. We have taken a simple sentence from a medical data set to understand the problem and where one can understand the terms '**cancer**' and '**chemotherapy**'.

**Text

Description automatically generated**

*Fig 1.1 Prototype of the Application*

**1.1 Motivation**

Named Entity Recognition (NER) is a vital step in medical information extraction, especially Electronic Health Records (EHRs). Proper extraction of medical entities such as disease and medications can automate the process of EHR coding as well as considerably improve the filtering of EHR resulting in better extraction of medical information.

NER systems are generally trained and evaluated on relatively small standard datasets. However, they are applied to real-world applications, and they are exposed to a different collection of texts, varying in topic, entity distribution, and text type. This mismatch between the internal structure and application can cause a drop in performance and consequently, unreliability.

The motivation behind choosing this topic was to create a prototype application “Doctor Shrewd” with high accuracy for EHR coding and filtering.

**1.2 Objective**

* Extract those tokens which have NOUN or PROPN as their PoS tag and find their frequency.
* Extracting dictionary of extracted diseases and potential treatments.
* Predict the treatment for the disease name.

1. **Literature Review**

Natural language refers to the way we, humans, communicate with each other. Namely, speech and text. Given the importance of this type of data, we must have methods to understand and reason about natural language, just like we do for other types of data. Natural Language Processing, or NLP for short, is broadly defined as the automatic manipulation of natural language, like speech and text, by software. The study of natural language processing has been around for more than 50 years and grew out of the field of linguistics with the rise of computers. **Named Entity Recognition (NER)** is an application of NLP. NER relies on NLP technology. Here,NLP identifies the Proper extraction of medical entities such as disease and medications can automate the process of EHR coding as well as considerably improve the filtering of EHR resulting in **better extraction of medical information**.

We have been given such a data set in which a lot of text is written related to the medical domain. There are a lot of diseases that can be mentioned in the entire dataset and their related treatments are also mentioned implicitly in the text, we saw in the example that the disease mentioned is cancer and its treatment can be identified as chemotherapy using the sentence.

However, note that it is not explicitly mentioned in the dataset about the diseases and their treatment, but we can build an algorithm to map the diseases and their respective treatment.

In this project, a NER tagger is proposed, covering mainly multiple entity types, medications and diseases. The model used is Conditional Random Fields (CRF) for sequence tagger.

1. **Design and Implementation**

**Step 1: Load and pre-process data**

### 

### **1.1 Load the data:** For text corpuswe have been given such a data set in which a lot of text is written related to the medical domain. There are a lot of diseases that can be mentioned in the entire dataset and their related treatments are also mentioned implicitly in the text

### 

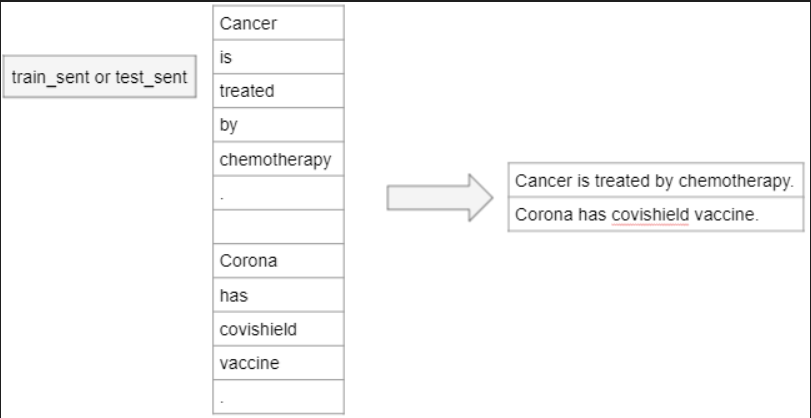
### **1.2 Pre-process the data:**

### The dataset provided is in the form of one word per line. Let's understand the format of data below:

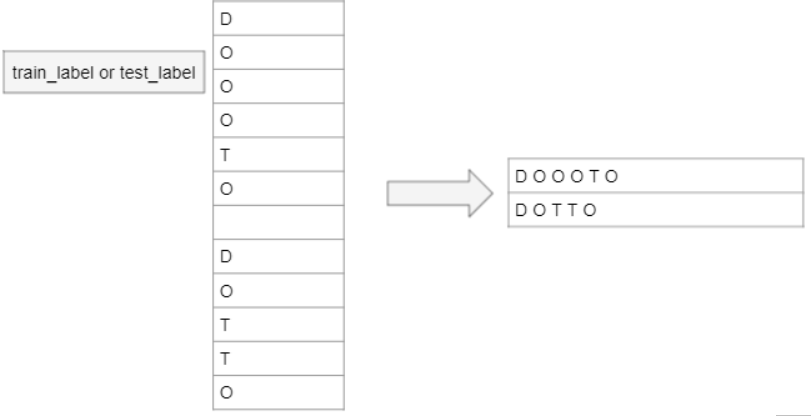
### - Suppose there are \*x\* words in a sentence, then there will be \*x\* continuous lines with one word in each line.

### - Further, the two sentences are separated by empty lines. The labels for the data follow the same format.

### We need to pre-process the data to recover the complete sentences and their labels.



**Fig 3.1 Data Set**

****

**Fig 3.1 Labelled Data**

### 

### **1.3 Split into train and test sets**

After tokenizing the data, we split it into training and test sets.

**1.4 Extract those tokens which have NOUN or PROPN as their PoS tag and find their frequency****.**

### **1.5 Pre-process the train and test data**

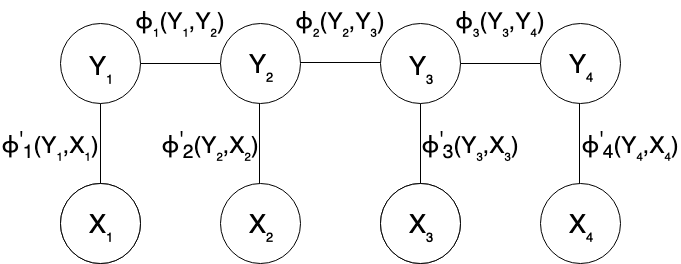
### Finally, we complete the pre-processing both for training and test sets.

## 

## **Step 2: Building the CRF model. Using max\_iterations as 200**

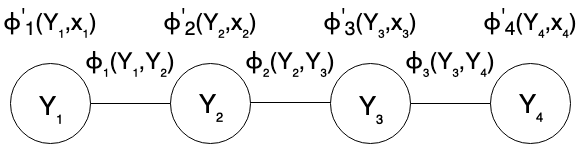
Conditional Random Field is a special case of Markov Random field wherein the graph satisfies the property : “When we condition the graph on X globally i.e. when the values of random variables in X is fixed or given, all the random variables in set Y follow the Markov property p(Yᵤ/X,Yᵥ, u≠v) = p(Yᵤ/X,Yₓ, Yᵤ~Yₓ), where Yᵤ~Yₓ signifies that Yᵤ and Yₓ are neighbours in the graph.” A variable’s neighbouring nodes or variables are also called the Markov Blanket of that variable.

One such graph that satisfies the above property is the chain-structured graph shared below: -



**Fig 3.1 Conditional Random Field structure**

Since CRF is a discriminative model i.e., it models the conditional probability P(Y/X) i.e., X is always given or observed. Therefore, the graph ultimately reduces to a simple chain.

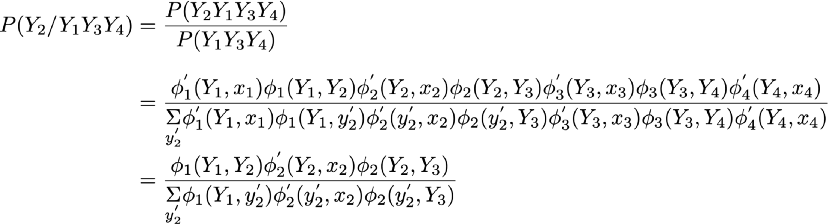


**Fig 3.1 CRF model, conditioned on X**

As we condition upon X and we are trying to find the corresponding Yᵢ for every Xᵢ, X and Y are also called the evidence and label variables respectively.

We can verify that the “factor reduced” CRF model shown above follows the Markov property as shown for variable Y₂ below. As shown, the conditional probability of Y₂ given all other variables finally depends only on its neighbouring nodes.

Variable Y₂ satisfying the Markov property, conditional depends only on neighbouring variables



**Fig 3.1 Variable Y₂ satisfying the Markov property, conditional depends only on neighbouring variables**

**Step 3: CRF model's prediction to prepare a record of diseases identified in the corpus and treatments used for the diseases.**

Create the logic to get all the predicted treatments (T) labels corresponding to each disease (D) label in the test dataset. Assumption: For each identified disease, one of the treatments is in the same sentence as the disease exists.

Table

Description automatically generated

***Fig3.1*** **Extracting a dictionary of all the predicted diseases from our test data and the corresponding treatments.**

1. **RESULT ANALYSIS**

Text

Description automatically generated

**Fig 4.1 Code Snippet for final Prediction**

Text

Description automatically generated

**Fig 4.2 Code Snippet for F1 score and its value**

1. **CONCLUSION**

We have built an algorithm to map the diseases and their respective treatment. Here, a NER tagger is proposed, covering mainly multiple entity types, medications and diseases. The model used is Conditional Random Fields (CRF) for sequence tagger. The train dataset is used to train the Continuous Random Field (CRF) model, and the test dataset is used to evaluate the built model

Using python as a language for the project, the basic steps of loading the data, pre-processing it, splitting it into training & test sets, followed by the pre-processing of training and test set was performed. Further steps involved developing CRF model, calculating the F1 score to evaluate the model on the test set. Finally, the model developed so far were combined to implement in an application (Doctor Shrewd) for searching treatment for given diseases.

In this project, NER identifies the Proper extraction of medical entities such as disease and medications can automate the process of EHR coding as well as considerably improve the filtering of EHR resulting in **better extraction of medical information**.

**References**

[1] Mohammed RAIS, Abdelmonaime LACHKAR, Said EL ALAOUI OUATIK(2014),A Comparative Study of Biomedical Named Entity Recognition Methods Based Machine Learning Approach.

[2] Ahmed Sultan Al-Hegami,Ameen Mohammed Farea Othman, Fuad Tarbosh Bagash,(2017). A Biomedical Named Entity Recognition Using Machine Learning Classifiers and Rich Feature Set.