NLP Methods for Adverse Drug Event Detection in Electronic Health Records

Nitika
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
nnitika@asu.edu

***Abstract*—Adverse Drug event is any kind of injury or adverse effect caused on administration of a drug and that drug may or may not be identified as the cause of the event. In medical world, a numerous data is available and generated such as electronic health records and discharge summaries etc. This data can be used as pharmacovigilance tool to identify the cause of adverse drug events and these events can be prevented from occurring in the future. This paper discusses the natural language processing techniques for bio medical text to extract relation between a drug and adverse drug events.**

***Index Terms*—ADE, EHRs, Drugs, NLP, BERT**

I. Introduction

Natural Language Processing is a branch of Artificial Intelligent that deals with how computers analyze human(natural) languages. Bio Medical Natural Language Processing deals with the analysis of clinical and medical text. Healthcare data is rapidly increasing and contains useful insights. As the volume of data increases which is largely in natural unstructured form. To draw useful insights from it, its imperative to develop robust natural language processing algorithms that can process and extract information efficiently and accurately. Bidirectional Encoder Representations from Transformers(BERT) is a natural language processing model developed by Google. It has achieved state-of-the-art performance for some of the challenging tasks of Natural Language Processing. The considerable amount of performance enhancement is attributed to its bidirectional nature and attention mechanism along with the use of pre-trained embeddings.

II. Related Work

Before the release of BERT by Google in 2018, Bi-LSTM-CRF (Bidirectional Long Short-Term Memory - Conditional Random Field) was a benchmark for Relation Extraction and Adverse Drug Event(ADE) detection from clinical text. Previous studies have also leveraged attention mechanism for NLP tasks such as relation extraction and joint entity recognition. [1]

III. Explanation of the solution

A. Task Specifications

In an EHR, entities such as Drugs, Reasons, and ADEs are annotated. The task is to identify the relationship between the following entities:

1) Drug-ADE
2) Drug-Reason
3) No relation

Relation extraction: It is a task to extract relations between entities in an Electronic Health Report. This project explores the relation extraction between following pair of entities:

1) Adverse Drug Events (ADEs) and Drugs
2) Reasons and Drugs.

B. Data Set

The data set used for this project is based on n2c2 2018 Track 2 task on Adverse Drug Events and Medication Extraction in EHRs. It contains discharge summaries of about 500 discharge summaries drawn from the MIMIC-III (Medical Information Mart for Intensive Care III) clinical care database. [2] The following entitles and relationships are annotated in the data set:

1) Entities: Entity type tags for Drug, ADE, and Reason.
2) Relations: Drug-Reason, Drug-ADEs

IV. Model

BERT is a Natural Language Processing model developed by Google. It provides text embeddings by capturing the contextual information from the text. The embeddings can then be used for downstream tasks and used by other machine learning models like Feed Forward deep learning network to perform classification and other tasks.

BERT is bidirectional model and it uses multi head attention mechanism that enables it to capture relation among tokens according to their positions in a text sequence. Masked Learning Models (MLM): MLM is used by BERT to train the model by masking parts of the input and then predicting the masked input. [3]

For this task, two BERT (Bidirectional Encoder Representations from Transformers) models for each relation type, Reason-Drug and ADE-Drug are used. BERT models with base configuration are used to perform binary classification to classify relations as:

1) ADE-Drug/ No relation
2) Reason-Drug/ No relation

A. ClinicalBERT

Since the process of generating embeddings is time consuming, therefore, to expedite the training, pretrained word

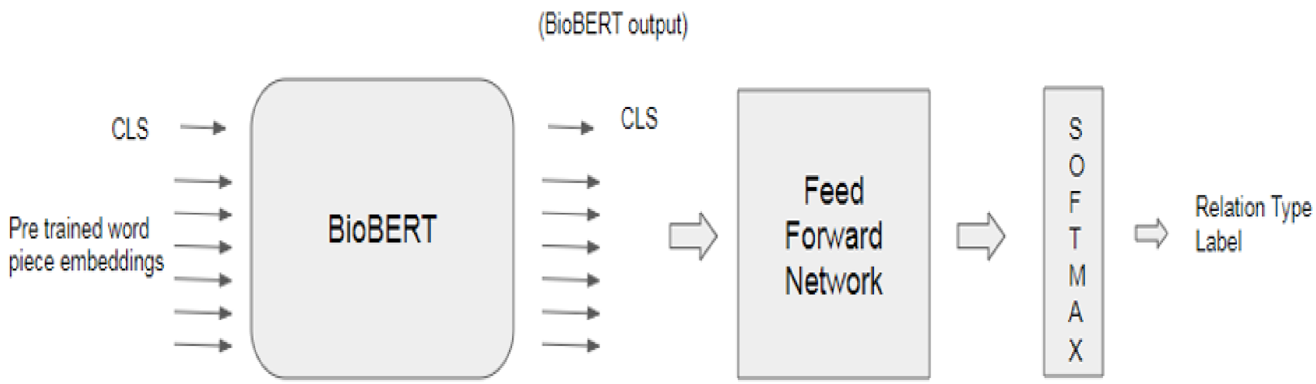

Fig. 1. NLP model for relation Extraction

embeddings are used as input to BERT. Model is then fine-tuned with the training data. For this project, pretrained embeddings (discharge summary embeddings) from Clinical-BERT are used. ClinicalBERT provides embeddings which are obtained by training BioBERT on about 2 million clinical notes from the MIMIC-III v1.4 database. [4]

B. *Fine Tuning*

Model is further fine-tuned with the provide training data set. *Input*: This model takes the Sequence of words as input. Internally it converts the sequence of words into word-pieces and the pre-trained embeddings of the word-piece tokens are finally used. *Output*: Model outputs the class label as 1 and 0. 1 denotes the relation exists and 0 indicates no relation.

C. *Span of Relations*

The span of the relation starts with the source entity and ends with the target entity and it can span multiple sentences. For simplicity sake and limitations of input size of BERT, the maximum length of the sequence is taken as 128 words. Sequence samples with length more than 128 words are removed from both training and testing set and performance is penalized for removing data samples. An example of a positive sample after preprocessing the data:

"@Reason$ treated by @Drug$"

V. Class Imbalance

Real life data is highly unbalanced i.e. there are more instances of one class than the other. Most machine learning algorithms does not perform well with class imbalance.

A. *Down sampling of negative examples:*

Data set can be balanced by down sampling the majority class i.e. negative samples in this case. For this task, down sampling is done by choosing sentences in which entities occur closer to each other i.e. sentences with shorter lengths over longer lengths.

The idea behind choosing this approach is to use the contextual information between the source and target entity pairs. Since in longer sequences, there can be multiple sentences between source and target entity which might not be relevant to the context as compared to shorter sentences where relevant contextual information can be captured. Two down sampling methods with ratio of 1 to 1 and 1 to 2 of positive to negative samples are used.

B. *Algorithm*

1) Create a list of all pairs of entities from both entity types and filter out the pair having relationship between them.
2) Create sequences from the pair starting with source entity and ending with target entity.
3) Sort the sequences by word count.
4) Take the top n sequences where n is equal to the number of positive relation pair from an EHR.

C. *Example 1*

Figure 2 shows a sequence of text from an EHR. The length of the sequence is considerably large with many tokens between source and target entity.

D. *Example 2*

Figure 3 shows a sequence of text from an EHR. The length of the sequence is relatively smaller than example 1 with very few entities between source and target entity.

The idea is based on the hypothesis that shorter length example captures better contextual information as compared to 1st example between reason and drug entity.

VI. Results

The results are illustrated in the following table using precision, recall and F- Score measures. The numbers are compared with state-of-the-art models from n2c2 2018 challenge on ADE detection and from Hong Guan and Dr. Murthy's paper

"@REASON with compression fracture but this did not seemto help with symptoms so was discontinued..# Leukocytosis: WBC >20 persistently in the MICU even afterbeing treated for infection. Since no new infection was foundthis was presumed [**12-26**] steroids and the leukocytosis improvedwith prednisone taper. WBC 12 on day of discharge.# Hyperglycemia: Patient is not known to be a diabetic and wasfelt [**12-26**] steroids, his sugars were controlled on sliding scaleinsulin in the hospital but he no longer had insulinrequirements as his prednisone was tapered..#. [**Last Name (un) **]: Cr 1.9 on [**3-18**] from 1.2 which improved to 1.4 on [**3-19**]with decreasing ACE and 500cc bolus. He should have repeatcreatinine and labs on [**3-22**] to ensure stability.# Guardianship: Guardianship paperwork was started in thehospital.Medications on Admission @DRUG"

Fig. 2. A sequence with large span

"@DRUG -patient may require additional nebs on top of his standingadvair though his respiratory status has been very stable,without @REASON"

Fig. 3. A sequence with small span

on Leveraging Contextual Information in Extracting Long Distance Relations from Clinical Notes. [5]

The results show that the model performed with high recall but low precision. The model predicted high number of true positives and low number of false negatives. P:N::1:1 – Positive to Negative samples ratio is 1 to 1 P:N::1:2 – Positive to Negative samples ratio is 1 to 2

TABLE I
ADE - DRUG RELATION

Metric	***P:N::1:1***	***P:N::1:2***	***N2C2 BEST***	***BERT+EDGE***
F1 Score	74.04	79.64	79.5	82.3
Recall	93.59	89.36		82.8
Precision	61.25	71.82		81.8

TABLE II
REASON - DRUG RELATION

Metric	***P:N::1:1***	***P:N::1:2***	***N2C2 BEST***	***BERT+EDGE***
F1 Score	74.74	78.65	75.8	83.2
Recall	96.27	93.89.36		85.8
Precision	61.07	67.67		80.7

VII. Reflection and Learnings

This project helped get acquainted with the concepts of Natural Language Processing in the field of bio-medical text-mining. The course and project forays into challenges faced in processing real patient data and its applications in the real world. The project gave an opportunity to read and analyze the relevant literature and state-of-the-art research papers and methods. This project provided with an opportunity to learn and implement new deep learning methods such as BERT. Along with concepts, through this project, new skills such as data mining, languages such as python, deep learning tools like jupyter notebook, google colab etc. are acquired.

VIII. Contribution

The project consists of two team members. Following is the details of the other team member who contributed on this project.

1) *Shubham Mittal* (CIDSE, Arizona State University), smitta21@asu.edu

Following is the description of my contribution towards this project:

1) *Ideation*: Definition of the scope of the project, preparation of the initial proposal.
2) *Literature survey*: Review of past research papers, Existing benchmarks
3) *Data Preprocessing*: Wrote code for forming sequence of text input from given entities.
4) *Model*: experimentation with different class ratio and sentence lengths to achieve high performance.
5) *Documentation*: Documented the findings and results of the project.

References

[1] Dandala B, Joopudi V, Devarakonda M. IBM Research System at MADE 2018: Detecting adverse drug events from electronic health records. In: International Workshop on Medication and Adverse Drug Event Detection. 2018. p. 39–47.

[2] https://portal.dbmi.hms.harvard.edu/projects/n2c2-t2/

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805

[4] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 72-78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

[5] Hong Guan, Murthy Devarakonda, "Leveraging Contextual Information in Extracting Long Distance Relations from Clinical Notes", unpublished.