Course Project Report

Comparative Study on Clinical Concept Extraction

Submitted By

Meghna Suraj Kashyap (181IT127)
Udbhav Bisarya (181IT150)
Vishwas Parekh (181IT252)

as part of the requirements of the course

Information Retrieval (IT458) [Jul - Nov 2021]

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology in Information Technology

under the guidance of

Dr. Sowmya Kamath S, Dept of IT, NITK Surathkal

undergone at

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

JUL-NOV 2021

DEPARTMENT OF INFORMATION TECHNOLOGY

National Institute of Technology Karnataka, Surathkal

C E R T I F I C A T E

This is to certify that the Course project Work Report entitled **"Comparative Study on Clinical Concept Extraction"** is submitted by the group mentioned below -

Details of Project Group

Name of the Student	Register No.	Signature with Date
Meghna Suraj Kashyap	181IT127	
Udbhav Bisarya	181IT150	
Vishwas Dipen Parekh	181IT252	

this report is a record of the work carried out by them as part of the course **Information Retrieval (IT458)** during the semester **Jul - Nov 2021**. It is accepted as the Course Project Report submission in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Information Technology.**

(Name and Signature of Course Instructor)
Dr. Sowmya Kamath S

D E C L A R A T I O N

We hereby declare that the project report entitled **"Comparative Study on Clinical Concept Extraction"** submitted by us for the course **Information Retrieval (IT458)** during the semester **Jul-Nov 2021**, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Information Technology at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Details of Project Group

Name of the Student	Register No.	Signature with Date
1. Meghna Suraj Kashyap	181IT127	
2. Udbhav Bisarya	181IT150	
3. Vishwas Dipen Parekh	181IT252	

Place: NITK, Surathkal
Date: 20/11/2021

Comparative Study on Clinical Concept Extraction

Meghna Suraj Kashyap[1], Udbhav Bisarya[2], Vishwas Parekh[3]

***Abstract*— Electronic health records are used to store patient treatment and diagnosis data, and have often been viewed of great importance in bringing about advances in the field of healthcare technology. Automatic extraction of concepts from these electronic health records proves to be a very essential step in changing the unstructured data within a health record or a clinical note into a form that has actionable and stuctured information. In this paper, we propose a comparison between different approaches that may be used to train a model for clinical concept extraction for the task of automatic annotation in the medical field, specifically for clinical problems, tests and treatments in clinical notes. We tested our proposed models on the I2B2 2010 challenge data set and compared the results with already existing approaches based on the F1 scores and other evaluation metrics.**

Keywords: **ELMO, BERT, NER**

I. Introduction

Electronic Health Records (commonly known as EHRs) are known to have great potential to improve clinical research and bring about advances in healthcare technology. A significant portion of clinical extraction is found as free text in the form of clinical notes, and we need to achieve the goal of extracting the "meaningful information" from this text, transforming the Electronic Health Record data into actionable knowledge.

Successfully being able to capture the key points and labels in the clinical records found in the form of free text can provide great insights into problems and possible solutions. It can provide insights into what treatments were successful for certain problems and what tests were able to accurately capture the underlying problems. It could also provide intra-patient links to understand how likely a patient is to a certain problem if they have experienced a similar one earlier.

The extraction of concepts from health records has been successful in the past few years with a lot of models based on NLP, but due to the heterogeneous nature of Health Record data meant that no one model has been successful globally. The particular class of NLP problems for extracting clinical concepts is Named Entity Recognition(NER). NER, also known as entity extraction is used to classify named entities present in text into annotations or pre-defined categories (in our case clinical concepts). It helps by adding semantic knowledge to our content and helps to understand the subject of any given text.
Another challenge towards the development of clinical Named Entity Recognition (NER) tools is related to access to a corpus of labeled data. In this paper, we have successfully trained an model that couples ELMo embeddings with LSTM and also a BERT model on the 2010 i2b2/VA data set and have done appropriate comparisons to understand what would be more optimal to use in clinical concept extraction.

II. Literature Survey

[1] This paper proposes a significant improvement to the previous state of the art machine learning models which are capable of interpreting patient records accurately albeit being more time and computationally intensive. The proposed approach relies on bi-directional LSTM models with general word neural word embeddings which are much less time consuming. The advantages this paper provides is that it offers significant improvement to state of the art models and this is seen to be simpler to implement on account of using off the shelf word embeddings. The paper also provides proof that Glove word embedding used here is not as efficient as using BERT or ELMo.

[2] This paper proposes a superior approach to patient records interpretation by utilizing the more efficient and informative contextual word embedding model ELMO than off the shelf embeddings. This new approach offers prominent and note worthy improvement in performance by 3.4% compared to state of the art models. Advantages of this paper includes superior results on account of using more sophisticated word embedding, ELMo rather than off the shelf like Glove. It brings best of both worlds and produces best results besting state of the art models by a considerable margin.

[3] This paper provides contributions by exploring the rather underutilized BERT word embedding to demonstrate its superiority over general word embeddings in both general clinical text and discharge summaries. Compares results of clinical BERT, general BERT and bio BERT with respect to de identification and explains the reasons for the reduced performance. This paper offers more insights into application of BERT which has seldom been utilized in the field.

III. Dataset

The data set used in our work is the data set provided by the 2010 i2b2/VA challenge for the task of training a concept extraction system. Due to the restrictions that have been introduced by the Institutional Review Board (IRB), only a portion of the entire data from the original data

set is available. The dataset released by them consists of clinical summaries from different sources, the sources are three different medical sites, namely Beth Israel Deaconess Medical Center, Partners Healthcare and the University of Pittsburgh Medical Center. The clinical concepts annotated in this corpus include three categories: problems, tests, and treatments.

IV. Methodology

We processed the i2b2 2010 dataset before using it to train our models for concept extraction, and then used two different approaches to training a model for the task. We used a BERT based approach and an ELMO and Bi-LSTM based approach to solve the NER (Named entity recognition) problem.

A. Preprocessing the Dataset

The zip files are extracted and we add the beth and partners folders to the training folder and the test dataset to the test folders. Once these are unzipped, the typos in these files are fixed and for ease of use the tokens and concepts are dumped into pkl files.

B. BERT based approach to solving NER

Once the preprocessing is done, the tokens and concepts are extracted from the pkl files (beth,patners and test). The training set, which composes of the beth and partners data, is then further split into training and validation set.

There is an existing clinical BERT model which has been trained on the MIMIC-III v1.4 dataset. This model, however, was not built for NER. We used the i2b2 dataset to fine tune this model for solving NER. To perform the fine-tuning we identified that the i2b2 dataset has 6 tags('problem', 'treatment', 'test', 'B-problem', 'B-treatment', 'B-test') and 1 outside tag (''). This is a similar format to the BIO Tag scheme. With a learning rate of 0.0001 and batch size of 10 the model is trained on the dataset and for each tag a corresponding f1-score, precision and recall is generated

C. ELMO and Bi-LSTM based approach to solving NER

Once the preprocessing is done, the tokens and concepts are extracted from the pkl files (beth,patners and test). The training set, which composes of the beth and partners data, is then further split into training and validation set.

ELMO based models give out vectors in the form of a matrix of the size (a,b,c) where a is the total number of sentences passed as input to the model, b is the length on the longest sentence, and c is 1024 which is the size of the embedding vector, therefore we padded each statement with a word so that all sentences are the same size, and the Bi-LSTM model could not have a standard input and output size.

We split the dataset into train (further divided into train and validate) and test, and then pass the train set into the ELMO model to get a vector as output of size (number of sentences,201,1024) (201 is because the length of the longest sentence was 201, and the rest of the sentences were then padded to be converted to the same length). We then take these ELMO embeddings and pass them through our Bi-LSTM model (we use a Bi-LSTM model instead of the traditional uni-directional LSTM, as in the case of concept extraction, we need to take into consideration not only the terms we have already encountered, but also the ones yet to come). The output contains a vector that is used to predict the label corresponding to each word within that sentence, there are 7 possible label categories ('problem', 'treatment', 'test', 'B-problem', 'B-treatment', 'B-test', ''), and each word in the corpus is mapped to one of the seven values, depending on the context behind its usage.

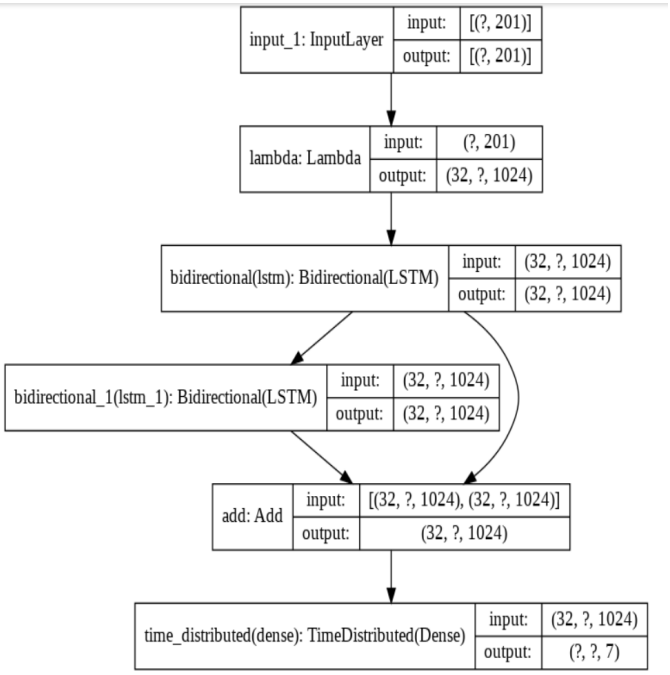

Fig. 1: Basic outline of the Bi-LSTM model for concept extraction

V. Results

A. BERT based approach to solving NER

	Level	F1-Score	Precision	Recall
0	problem	0.899204	0.891408	0.907137
1	treatment	0.860257	0.861451	0.859066
2	test	0.873740	0.876428	0.871068
3	B-problem	0.913256	0.914602	0.911914
4	B-treatment	0.896908	0.900054	0.893783
5	B-test	0.897152	0.884534	0.910136
0	AVG_MICRO	0.893413	NaN	NaN
0	AVG_MICRO	0.890086	NaN	NaN

Fig 2: BERT Results

B. ELMO and LSTM based approach to solving NER

	precision	recall	f1-score	support
test	0.78	.76	.77	201
B-problem	.87	.75	.81	668
treatment	.59	.76	.67	221
problem	.72	.70	.71	434
B-test	.89	.73	.80	537
B-treatment	.79	.81	.80	485
micro avg	.79	.75	.77	2546
macro avg	.77	.75	.76	2546
weighted avg	.80	.75	.77	2546

Fig 3: ELMO and Bi-LSTM Results

C. Comparing results with benchmark

Methods	precision	recall	f1-score
ELMo(General) + BiLSTM-CRF	0.8326	0.8184	0.8254
ELMo(Clinical) + BiLSTM-CRF	0.8934	0.8787	0.8860
ELMo(General) + BiLSTM	0.79	0.75	0.77
BERT			0.8900

Table 1: Comparing our proposed models with results with [2]

VI. Conclusion

In natural language processing, contextual word embedding approaches have been optimistic. In this paper we have used ELMo along with Bi-LSTM and BERT models which is trained on 2010 i2b2/VA dataset. The ELMo along with Bi-LSTM model has an f1-score of 0.77 and the BERT model has an f-score of 0.890086. Comparing with the methods used in [2], the ELMO(General) and simple Bi-LSTM performs poorly when compared with the BiLSTM-CRF based models, while the BERT model used by us outperforms the models proposed by them. The BERT model is seen to perform better and is more optimal to use in clinical concept extraction. This paper provides extensive comparison between models and shows the effectiveness of NER tasks to provide meaningful results for clinical diagnosis.

REFERENCES

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W., Jin, D., Naumann, T., and McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings. *CoRR*, abs/1904.03323.

Chalapathy, R., Borzeshi, E. Z., and Piccardi, M. (2016). Bidirectional LSTM-CRF for clinical concept extraction. *CoRR*, abs/1610.05858.

Zhu, H., Paschalidis, I. C., and Tahmasebi, A. (2018). Clinical concept extraction with contextual word embedding. *CoRR*, abs/1810.10566.

(Chalapathy et al., 2016) (Alsentzer et al., 2019) (Zhu et al., 2018)