**1. Mathu T[1], 2. Kumudha RAIMOND[1] 3. Deepakanmani S[2]**

Karunya Institute of Technology and Sciences (1), Sri Krishna College of Engineering and Technology (2)
ORCID: 1. 0000-0003-1674-3418; 2. 0000-0001-8680-8390; 3. 0000-0001-8461-3792;

doi:10.15199/48.2023.08.18

# A hybrid drug named entity recognition framework for real time pubmed data using deep learning and text summarization techniques

***Abstract**. Drug Named Entity Recognition (DNER) becomes indispensable for various medical relation extraction systems. Existing deep learning systems rely on the benchmark data for training as well as testing the model. However, it is very important to test on the real time data. In this research, we propose a hybrid DNER framework where we incorporate text summarization on real time data to create the test dataset. We have experimented with various text summarization techniques and found SciBERT model to give better results than other techniques.*

***Streszczenie.** Rozpoznawanie jednostek o nazwie leku (DNER) staje się nieodzowny dla innych systemów ekstrakcji relacji medycznych. Istniejące systemy głębokiego uczenia się opierają się na danych porównawczych zarówno podczas szkolenia, jak i testowania modelu. Jednak bardzo ważne jest, aby testować dane w czasie rzeczywistym. W tym badaniu proponujemy hybrydową strukturę DNER, w której uwzględniamy podsumowanie tekstu na danych w czasie rzeczywistym w celu utworzenia zestawu danych testowych. Eksperymentowaliśmy z różnymi technikami podsumowania tekstu i stwierdziliśmy, że model BERT daje lepsze wyniki niż inne techniki. (**Hybrydowa struktura rozpoznawania jednostek o nazwie lek dla publikowanych danych w czasie rzeczywistym przy użyciu technik głębokiego uczenia się i streszczania tekstu**)*

**Keywords:** Drug Named Entity Recognition, deep learning, text summarization, BERT
**Słowa kluczowe:** Rozpoznawanie jednostek o nazwie leku, głęboka nauka, podsumowanie tekstu, BERT

## Introduction

The volume of scholarly articles in the biomedical field has significantly increased in recent years. Most of this literature can be found and easily accessed in electronic form. This unstructured text can provide numerous valuable information for researchers. Biomedical text mining techniques need to be applied to extract this useful information. Information Extraction (IE), a Natural Language Processing (NLP) task, analyses documents written in natural language with the goal of extracting structured and practical information, such as named entities and semantic relationship between them [1]. The prominent entities present in biomedical text are drug, disease, protein, cell, genes, chemical compounds, etc. Named Entity Recognition (NER) would be the most basic step in any IE process. The method of identifying the drug entity from the unstructured textual data is called Drug Named Entity Recognition (DNER) [2]. The drug entity is significant in the medical extraction systems such as Drug-Drug Interaction (DDI) [3] and Adverse Drug Reaction (ADR) [4]. The extensive analysis required for such research demands the researchers to read and process thousands of documents. The existing systems develop several state-of-the-art machine learning (ML) and deep learning (DL) models for DNER using various training and test datasets available. However, the real time data available in the sources such as PubMed in the form of journal articles are huge and lengthy and practically impossible to read through the entire document. Text Summarization (TS) comes as a solution to overcome this problem [5]. TS condenses the size of the research articles to make it easier for users to access and analyze essential source materials. There are two categories of Text Summarization, namely, Extractive Text Summarization (ETS) and Abstractive Text Summarization (ATS). ETS produces an extractive summary that are direct excerpts from the input text. This summary would be a regressive conversion of the original text into the summary text using substance minimization or generalisation based on what is crucial in the original document. This is a promising method to create a crisp and elegant summary of huge and long documents while retaining the core concepts and significance. Once the documents are summarized, the summarized text can be processed and tokenized. The tokenized dataset can be used as the test dataset for any ML or DL model to recognize the drug entities. Unlike ML, DL techniques doesn't use handcrafted features and hence proved to be the state-of-the-art for any NER models. The development of DL methods used in NLP has enabled it for biomedical NER to leverage TM frameworks.

The main contributions of this paper are a) Proposing a framework to evaluate DNER model with real-time test dataset b) Implementation and comparison of various BERT based ETS.

## Methodology: Framework for evaluating DNER model with real-time test dataset

The proposed DNER framework consists of the following phases: Input Phase, Text Summarization Phase, Training Phase, Testing Phase and Output Phase as shown in the Fig.1.

### 1. Input Phase

Two kinds of input need to be collected and processed for this framework. One set of data is to train the DL based DNER model and another set of data is to test the trained model. To train the model, various drug corpuses may be used. DDI 2013 corpus [6] contains abstracts from MedLine and DrugBank databases. It can be preprocessed and converted into tokens and tags in a csv format. The next set of datasets can be taken from PubMed, a huge database comprising more than 34 million citations for biomedical literature from various sources including Journals and Online books. Real-world scientific research greatly benefits from analyzing the enormous and continually expanding corpus of scholarly text data. The input documents can be found from the database using specific keyword / phrase search based on the research to be done. For instance, to do research on the diabetes disease and the drugs used for it, the database can be searched with a phrase "diabetes and drugs". This would result in more than 60,000 research publications. The documents resulted from the keyword search could be initially reduced by applying filters such as 'Results by Year', 'Text Availability', 'Article Type' and 'Publication Date' available on PubMed. The initial filtration done in the database would reduce the number of results. It would be helpful in doing the research specifically like "drugs that are commonly used for diabetes in the past 5 years". Also, the input could be a single document or the abstracts of multiple documents.

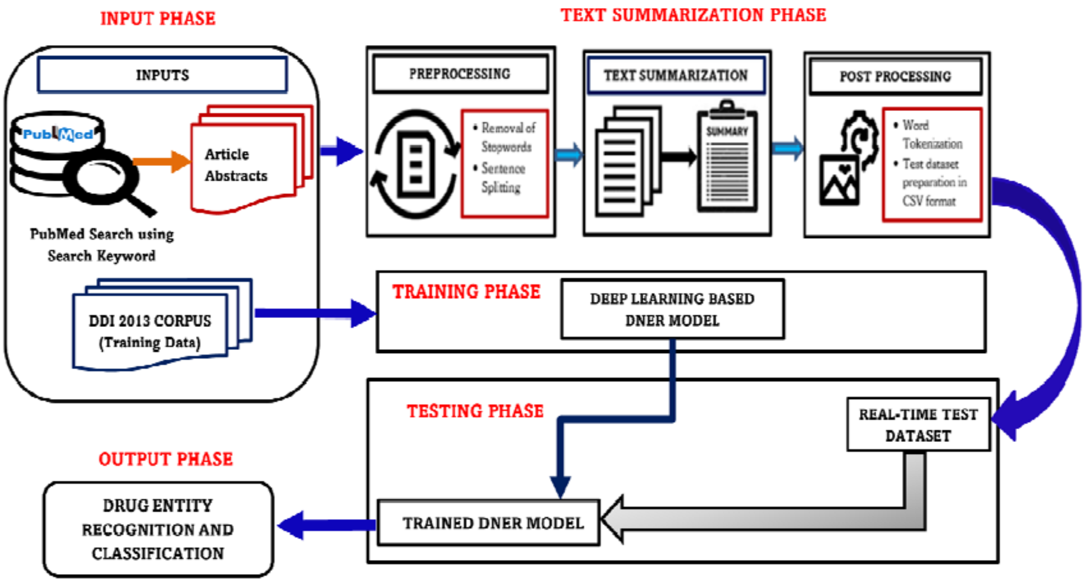

Fig.1.

Framework of the hybrid DNER model

## 2. Text Summarization Phase

This phase involves a) Preprocessing b) Extractive Text Summarization (ETS) and c) Post processing. It has been attempted to use statistical and machine learning techniques such as word relevance and also on the basic principle of the score of TF-IDF to solve for automated text summarization. Deep-learning techniques have changed the emphasis from manually building the features towards a more data-driven method, in which features are extracted and utilised to categorise phrases whether to include in the summary or not. [7]

a) Preprocessing: The input documents collected from the database needs to be preprocessed. The pre-processing includes sentence tokenization and stop word removal of the input articles.

b) Extractive Text Summarization: In biomedical data, it is essential to extract the sentences containing the key biomedical entities. The ETS method entails extracting the most significant sentences from the documents. The summary is then created by combining all the important sentences. Hence, every line and word of the summary is taken from the original document that is being summarized. There are several TS algorithms in the literature which includes Latent Semantic Analysis (LSA), Luhn, SumBasic, KLSum, LexRank and TextRank [8]. TextRank algorithm proves to be the best among these algorithms. It is a graph-based algorithm where it uses co-occurrence to create the graph with each sentence as the nodes of the graph. The top-n rank sentences is then extracted from the resultant graph by using PageRank algorithm. These sentences are finally used to form the summary of the input documents. However, since most of these algorithms uses the concept of TF-IDF [9], it would not be a better option for biomedical documents. Because, if we consider, for example, an abstract about "drugs used for diabetes", drug entities relevant to it might not appear frequently in the text. Hence these algorithms may not extract the right sentences in the summary. But, DL based BERT model [10] for ETS, implements by first embedding the sentences and then by applying a clustering algorithm, the sentences are extracted that are nearest to the cluster's centroids. Hence the sentences extracted for summary would be more relevant.

Methods used within BERT:

BERT is a pre-trained model, hence there is no training of the model required in the code. We have used many variations provided within BERT. The first variation is using hidden layers as the embedding output. These hidden layers cluster sentences based on their importance to the overall text. There is another variable of BERT called sentence BERT (SBERT). The improvement of SBERT when compared with BERT is that the model uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. This reduces the effort for finding the most similar pair from 65 hours with BERT / RoBERTa to about 5 seconds with SBERT, while maintaining the accuracy from BERT. A key feature that we used to identify the optimal number of sentences present based on the input text is the concept of "ELBOW". it is based on the technique used in cluster analysis. In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. There is another subset of BERT called Sci-BERT where the BERT model has been trained on thousands of research papers. This is specifically targeted at research papers. The results of all the models are detailed in the results section.

c) Post-Processing and Real-Time Test Dataset

Word tokenization is then applied to the summary to form tokens. While tokenizing the biomedical summary, care must be taken so that the original data is not lost. For instance, the word "catechol-o-methyl" is a single word and the hyphen between them needs to be retained to identify the right entities. The tokens are then stored in a csv format for further processing. In this framework, we can use the tokens formed from the real-time data which is summarized using text summarization.

## 3. Training Phase of DNER Model with Deep Learning Techniques

The DL model becomes the state-of-the-art for NER tasks. DL employs multiple layers of artificial neural networks to identify the named entities. DL is more efficient in identifying hidden features when compared to conventional techniques. Long short-term memory (LSTM) is a popular DL model that aids in preserving the long-range dependency specifically when working with the sequential text. While using LSTM, any word or character embedding model is used in converting each word to a vector. For better prediction, it is necessary to take the context of the word. Reading the text both in forward and reverse direction using Bidirectional LSTM (Bi-LSTM) would help to improve capturing the context better. In [11], the authors have used Recurrent Neural Network (RNN) model with Bi-LSTM and Bi-LSTM CRF along a with a specific word embedding model. BiLSTM-CRF has attained the improved performance than the previous models. Again in [12], an improved F1-score is obtained with a DL model of LSTM-CRF with word embedding. In word embedding models like Glove, Word2vec and FastText, same vector is used for all the mentions of a particular entity. This becomes a limitation in identifying the context or the semantic feature of the words correctly. In [13], we have worked on a DL model consisting of a stacked bi-LSTM and a residual LSTM along with a sentence embedding model instead of word embedding model. Also, we have used the Embedding from Language Model (ELMO) [14] sentence embedding model which acts on the whole sentence and hence able to identify the context correctly. The primary benefit of this model is that it can create vectors to the words which is not seen during training. This model is compared with LSTM-CRF [15], LIU [16] and WBI [17] and found improvement in terms of micro-average F1-score as 11.17, 8.8 and 17.64 respectively. In addition to that, the 83.89% of 2-gram and 76.67% of 3-gram entities are recognized correctly.

## 4. Testing phase of DNER model with Summarized real-time test data set

The real-time test dataset prepared could be given to the trained DNER model to recognize the drug entities. The trained model works on the test dataset to recognize the drug entities.

## 5. Output Phase

This phase includes the identification of drug entities with four categories of pharmacological entity classification such as group, brand, drug and drug_n [18]. Group defines the chemical relationship between drugs. Brand is the name of the chemical substance given by a pharmaceutical company which developed it initially. Drug represents a chemical substance which is approved to be utilized for humans to cure or diagnose a disease. It may also be used for the prevention of the disease. Finally, drug_n is any chemical substance which is unauthorized to be used for human beings.

### Performance Metrics

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is the main performance metric used to know how well the summary has been constructed. There are several types of Rouge namely: ROUGE-N where N represents the number of grams as in (1), ROUGE-L which measures the longest common subsequence (LCS) between our model output and reference, wherein a longer shared sequence would indicate more similarity between the two sequences and ROUGE-S which uses the skip-gram metric allows us to search for consecutive words from the reference text, that appear in the model output but are separated by one-or-more other words.

(1) Rouge-n = no. of n-grams in model and reference / no. of n-grams in reference

(2) Precision = TP / (TP+FP)

(3) Recall = TP / (TP+FN)

where TP is True Positive, FP is False Positive and FN is False Negative

(4) F1-score = 2. (Precision.Recall) / Precision + Recall

We have experimented with rouge-1 and rouge-2 measures and the results are discussed in the next section.

### Results and Discussion

We have implemented the text summarization model and NER model in Google Colab. The real time input data is taken from Pubmed. Initially we have collected 1000 abstracts with search keyword ”parkinson disease and drugs” from pubmed. They are preprocessed and then summarized using various models. The results obtained for extractive text summarization using each model is given in the table below. Rouge-1 and Rouge-2 performance metrics are used to compare the precision, recall and F1-values using the formula given in (2), (3) and (4) respectively.

Table 1. Rouge-1 Precision, Recall and F-score values (in %) of various models

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| BERT | 100 | 24.7 | 39.6 |
| BERT with a hidden layer | 98.7 | 36.69 | 53.4 |
| SBERT | 98.36 | 43.16 | 60 |
| SciBERT | 100 | 42.44 | 59.59 |

Table 2. Rouge-2 Precision, Recall and F-score values (in %) of various models

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| BERT | 95.51 | 18.16 | 39.6 |
| BERT with a hidden layer | 95.16 | 30.56 | 46.17 |
| SBERT | 96 | 37.3 | 53.73 |
| SciBERT | 98.5 | 34.19 | 50.76 |

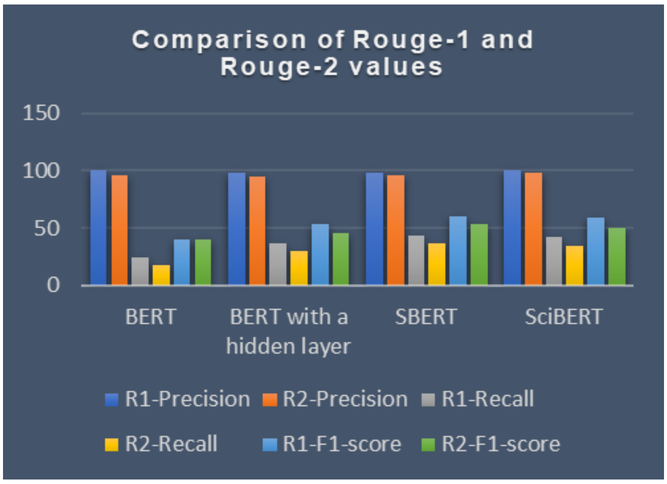

Fig.2. Comparison of Rouge-1 and Rouge-2 w.r.t precision, recall and F1-score with various BERT models

The results from Table 1. and Table 2. shows that the Rouge-1 values for SBERT model gives a better summarization model when compared with others. SciBERT also gives values closer to SBERT. However, when the actual results are analyzed by the experts, the

SciBERT model is better among all other BERT models for biomedical based text summarization process. In continuation to that we have implemented the DNER model as discussed in the training phase of the methodology. It not only improves the classification process but also decreases the time of processing.

## Conclusion

In this paper, we have proposed a hybrid drug named entity recognition framework incorporating the extractive text summarization technique. The various models of BERT was implemented for biomedical text summarization and found SciBERT to be the better model. The perfomance of text summarization is calculated using rouge-1 and rouge-2 values and compared. As a future scope, this work can also be implemented on full text research papers. Various DL models can be implemented for DNER classification.

***Authors**: Mrs. T Mathu, Dr. Kumudha Raimond, Department of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India. Dr. DeepaKanmani S, Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore, India.*
*E-mail: mathu@karunya.edu, kraimond@karunya.edu, deepakanmanis@skcet.ac.in*

## REFERENCES

[1] Ghoulam A., Barigou F., Belalem G., Information extraction in the medical domain, *Journal of Information Technology Research (JITR).*, *8*(2015), No.2, 1-15.

[2] Korkontzelos I., Piliouras D., Dowsey A. W., Ananiadou S., Boosting drug named entity recognition using an aggregate classifier, *Artificial intelligence in medicine.*, *65*(2015), No. 2, 145-153.

[3] Cheng L., Chang T. K., Wong, H., Drug-Drug Interactions With a Pharmacokinetic Basis, *Compr. Pharmacol.*, (2022), 698–715.

[4] Kant A., Bilmen J., Hopkins P. M. Adverse drug reactions, *Pharmacology and Physiology for Anesthesia.,* (2019),130-143.

[5] Allahyari M., Pouriyeh S., Assefi M., Safaei S., Trippe E. D., Gutierrez J. B., Kochut K., Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268.*, (2017).

[6] Herrero-Zazo M., Segura-Bedmar I., Martínez P., Declerck T., The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. Journal of biomedical informatics, 46(2013) No.5, 914-920.

[7] Bhargava R., Sharma Y., Deep extractive text summarization, *Procedia Computer Science.,* 167 (2020), 138-146.

[8] Zhang M., Li X., Yue S., Yang L., An empirical study of TextRank for keyword extraction, *IEEE Access*, 8(2020), 178849-178858.

[9] Qaiser S., Ali R., Text mining: use of TF-IDF to examine the relevance of words to documents, *International Journal of Computer Applications*, *181*(2018) No.1, 25-29.

[10] Miller D., Leveraging BERT for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*. (2019)

[11] Unanue I. J., Borzeshi E. Z., Piccardi M., Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition, *Journal of biomedical informatics*, *76*(2017), 102-109.

[12] Habibi M., Weber L., Neves M., Wiegandt D. L., Leser U., Deep learning with word embeddings improves biomedical named entity recognition, *Bioinformatics*, *33*(2017) No.14, i37-i48.

[13] Mathu T., Raimond K., A novel deep learning architecture for drug named entity recognition. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *19*(2021) No.6, 1884-1891.

[14] Peters M. E., Neumann M., Iyyer, M., Gardner, M., Clark C., Lee K., Zettlemoyer L. (1802) Deep contextualized word representations. CoRR abs/1802.05365 (2018). arXiv preprint arXiv:1802.05365.

[15] Zeng D., Sun C., Lin L., Liu B., LSTM-CRF for drug-named entity recognition, Entropy, 19(2017), No. 6, 283.

[16] Liu S., Tang, B., Chen, Q., & Wang, X. Effects of semantic features on machine learning-based drug name recognition systems: word embeddings vs. manually constructed dictionaries. *Information*, 6(2015), No. 4, 848-865.

[17] Rocktäschel T., Huber T., Weidlich M., Leser U., WBI-NER: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs, In Second Joint Conference on Lexical and Computational Semantics, Proceedings of the Seventh International Workshop on Semantic Evaluation, 2(2013), 356-363.

[18] Abacha A. B., Chowdhury M. F. M., Karanasiou A., Mrabet Y., Lavelli A., Zweigenbaum P., Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug–drug interaction extraction and classification, Journal of biomedical informatics, 58(2015), 122-132.