Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-021-03078-z

ORIGINAL RESEARCH

Named entity recognition on bio-medical literature documents using hybrid based approach

R. Ramachandran[1] · **K. Arutchelvan**[1]

Received: 22 November 2020 / Accepted: 2 March 2021

Abstract
There have been many changes in the medical field due to technological advances. The progression in technologies provides lot of opportunities to extract valuable insights from huge amount of unstructured data. The literature documents published by the researchers in medical domain consists enormous amount of knowledge. Many organizations are involving in retrieving the hidden information from the literature documents. Extracting the drug names, diseases, symptoms, route of administration, species and dosage forms from the textual document is an easy task due to the innovation of technologies in the Natural Language Processing. In this article, a new hybrid based approach is proposed to identify named entity from the medical literature documents. New dictionary has been built for route of administration, dosage forms and symptoms to annotate the entities in the medical documents. The annotated entities are trained by the blank Spacy machine learning model. The trained model provide a decent accuracy when compared with the existing model. The hybrid model is validated with the dictionary and human (optional)to calculate the confusion matrix. It is able to identify more entities than the prevailing model. The average F1 score for five entities of the proposed hybrid based approach 73.79%.

Keywords Natural Language Processing · Dictionary Based Approach · Named Entity Recognition · Machine Learning · Transfer Learning

1 Introduction

The emerging technologies in the digitalized era lead to the advancement of *Artificial Intelligence* (AI). Machine learning one of the prominent field of AI which reduces man power and upsurges insights from the huge amount of data. The role of AI in various domain such as education sector, financial management, transportation, healthcare management, etc. produces lot of opportunities to the AI researchers and skilled developers. Healthcare management is the part of Life science field which has emerging with lot of advancement with the AI technologies (Shah A.M. et.al. 2020). In this Covid-19 a pandemic situation, AI has being played a major role to find the presence of disease. Researchers has found the disease by training datasets available by using machine learning model and able to predict the presence of Covid-19. Still it is necessary to increase the accuracy of the machine learning model.

On the other hand, huge amount of articles related to the life science has being published in the popular databases such as PubMed, WebMD, etc. These articles are having lot of hidden information. Many organization and drug manufacturers has invested to extract the hidden information from the massive amount of data. Pharmacovigilance is one of the major area, where, it has to monitor the label document of the drug before its release in to the market. Extracting the entities such as diseases, chemical compound, active ingredients, gender, symptoms, dosage levels, dosage forms, route of administration, date, location, species, adverse reaction, etc. are difficult task. The massive amount of literature documents available over the internet are in unstructured format (Missen et.al. 2020) For example, electronic health records, clinical trials are has different format based on the countries regulations. Identifying the above said entities from the unstructured document is a challenging tasks for the researchers and developers. Generating the toxicology report immensely depends on these entities, to extract the case studies and other information. These entity extraction

✉ R. Ramachandran
ramachandranr.au@gmail.com

[1] Department of Computer and Information Science, Annamalai University, Tamil Nadu, Chidambaram, India

Published online: 11 March 2021

helps numerous researchers and report generators to reduce the manual time taking for producing the reports for clinical study, electronic health records, toxicology reports, etc. Most of the documents are in unstructured format and difficult to access the information. Several approaches such as dictionary based, machine learning based and hybrid based approached are proposed to extract these entities from the huge amount of unstructured data. Dictionary based method is more cost effective when compared to the machine learning based approach. Maintaining and updating the corpora leads to huge hardware and storage expenses. Machine learning (Jing Li et. al., 2018) based method which is further transformed into Transfer Learning (TL) produces more accuracy.

Building a TL model is being easy by the recent progression of the database technologies which ease to construct a huge amount of corpora. The corpora can be built from the information available in the pharmaceutical web pages. These information can be used for the various Natural Language Processing (NLP) applications to get the wisdom. NLP undertaking these challenges and provide the valuable insights. This emerging technology needs life science dictionaries and working machine learning model to predict the named entities. This paper has been concentrated on Named Entities Recognition (^{NER}) on the literature documents available over the internet. Hybrid based approach has been proposed to identify the entities. Diseases, chemical compound, symptoms, dosage forms and route of administration. Separate dictionaries has been built for the above mentioned entities. The unstructured documents are converted into a structure format and the named entities are matched with the string in dictionaries. The matched words are annotated with the start and end position of the characters. The annotated sentences are trained with the SciSpacy model and custom spacy model was built. The model was evaluated with the dictionary and confusion matrix is calculated.

This research article is organized as follows. Background study about the existing NER and NLP concepts are described in the section 2. Motivation of this research work is discussed in section 3. Proposed hybrid NER approach is discussed in the section 4 with detailed architecture and algorithm. Results and discussions are presented in section 5. The article is concluded in the section 6.

2 Background Study

Rapid advancement in technologies generates more number of resources and research articles in life science domain which has immense valuable information and becoming challengeable for examining and retrieving the insights (Galea D. et. al., 2018). According to the researchers (Rawassizadeh R et.al. 2015 and Thorne S 2000), refinement of unstructured data that has extracted from the various resources would provide more knowledge and production. In life science field, recognizing and extracting the named entities such as chemical, symptoms, ingredients, diseases, genes, dosage level, dosage forms, active substances, etc. from the unstructured texts is being carried out by the Natural Language Processing (NLP). Among may tasks, NER is one of the basic task in the NLP.In textual information, for retrieving the insights, identifying the important terms by means of the relationship among the words are more significant. To identify the expressive phrases or word in a field, categorizing the similar subject or object under an entity, is known as Named Entity Recognition (NER)(Nadeau Det.al. 2007; Grishman R et.al. 1996; Sudhakaran Gajendran et.al. 2020 and Guihua Wen et.al. 2020).

The authors (Guillaume Lample et.al. 2016) discussed about the Pharmacovigilance (PV), one of the prominent area in life science domain, is described as "the science and exercises identifying with the discovery, appraisal, comprehension and counteraction of unfriendly impacts or some other medication related issue." Pharmacological organizations regularly need to comprehend the environments and pre- environments below which a medication may have an unfavorable response on a patient. This would help in exploration and investigations of the medications and furthermore decrease or forestall dangers of any mischief to the patient. To meet out these properties, NER plays a major role to extract the named entities. Co-event of infection and synthetics in an (Electronic Medical Record) EMR of a sick person is helpful in readings and exploration for most drug organizations. Notwithstanding, EMRs are free-structured information and extra preparing is needed to extricate organized data, for example, named substances of intrigue. Such extraction can prompt noteworthy reserve funds of physical work and limiting the time taken to get another medication to advertise.

The authors (Giorgi JM et. al., 2018) discussed that, NER is playing a significant part in text mining to extract the relation among the objects by means of the annotations. Henceforth, to analyze the literature document published in the life science domain, NLP plays a significant role with effective and efficient approaches to extract the meaningful information (Wang X et.al. 2018). Annotating these information manually would consume lot of man-power and more cost effective. The process would be more complex and time consuming to extract the insights from the huge amount of data (Snow R et.al. 2008).

The authors (Rau LF 1991; Cho H et.al. 2017; Zhu Q et.al. 2017 and Leser U et.al. 2005) have discussed about the rule-based NER approaches on the textual information. Before the innovation of machine learning in the life science domain, dictionary-based approach have been used. The dictionaries are large collection of predefined entities.

The dictionary based approach are run by the rule based NER. Most of the rules are manually generated. The dictionaries are maintained separately for diseases, chemicals, genes, etc (Rocktaschel T et.al. 2012). Unlike the NER in other domain, life science domain is more complex because of the following concepts:

i. New inventions are published day-by-day
ii. Enormous amount of synonyms
iii. Huge numbers of abbreviations
iv. Lengthy phrases
v. Mixed combinations of punctuation, letters and symbols.

Applying rule based NER on the life science domain become more complex because of the above mentioned issues .

In the recent years, the applications of machine learning hybrid with different approaches produce more accuracy (Oudah M et.al. 2012). Deep learning which is a subset of ML, is used in the life science domain for various purposes. Predicting diseases, gene patterns are being carried out by the deep learning approaches . The interventions of statistical word padding and deep learning in NER developed various approaches. BiLSTM (Bi-directional Long Short Term Memory) is one of the best approach recommended using the deep learning. ChemSpot, a NER tool to identify chemical compounds in the documents is developed by using the BiLSTM and Conditional Random Filed (CRF) concepts (Lample G et.al. 2016; Habibi M et.al. 2017 and Leaman R et.al. 2013).

Apart from the deep learning, the following supervised machine learning approaches are widely used in most of research works (Chieu HL et.al. 2002; Settles B 2004; Isozaki H et.al. 2002 and Kazama JI et.al. 2002). The approaches are Support Vector Machine (SVM), Hidden Markov Models (HMM), Maximum Extropy (ME) and Conditional Random Fields (CRF). The most of ML tools were developed to identify entities automatically, yet they did not produce 100% accuracy and needs domain expert to validate and retrain (Ma X et.al. 2016; Gridach M 2017 and Zhao Z et.al. 2017). In other domain such as social media domain the NER for contextual information has been implemented. In biomedical domain, the contextual information needs a significant enhancements (Schnall A et.al. 2014). The author (Melamud et. al., 2016) presented the application of contextual information in the life science domain. They had achieved the tasks by using the word to vector and word embedding concepts.

On the other hand, Transfer Learning (TL) is an emerging field of ML, which is proposed to work a job on the target dataset by using the information learned from the source dataset (Li Q 2012; Pan S.J. et.al. 2010 and Weiss K et.al. 2016). Additionally, TL decreases the computation time of the model training, also, the amount of labeled data, and produce high performance. The TL uses silver standard corpus and gold standard corpus for large lower quality corpus and smaller standard quality corpus respectively (Rebholz-Schuhmann D et.al. 2010). BioNER (Devlin J et.al. 2018) is one of the prominent example of application of TL in biomedical NER. They have showed the higher performacnce and improvement in the NER accuracy.They used the concept BiLSTM-CRF method to provide better accuracy. This is one of the most powerful NLP resource to identify the named entities. BERT is developed using the deep learning concepts and trained with 2.5 billion words .

From the detailed background study, this research work is mainly focused to increase the number of entities. Attempted to identify symptoms, dosage forms and route of administration entities by built new dictionaries. Furthermore, hybrid the ML concept along with dictionary based method to train the model and achieve better accuracy. Section 3 presents the motivation of this research work.

3 Motivation

Current progressions in the NLP have presented significant enhancements in various NLP tasks by using the TL. BERT is one of the reliable resources that encouraging the TL. The models have been developed to supervised themselves and predict the next token in the sentences. Based on the domain, it can be easily trained and supervised to categorize the entities. TL reduces the training time and produces better accuracy. Most of the applications such as chatbot, automated forums, etc. are using the transfer learning concepts to train itself. In NLP, the traditional annotation process in very time consuming. It is the basic process of NLP.

This research work aims the following:

i. to annotate the sentence faster
ii. to implement the transfer learning
iii. to validate the output using the dictionary based or human-in-the-loop

4 Materials and Methods

From the background study it clearly indicates very less amount of work has been carried out to extract the route of administration entities, symptoms and dosage forms. In this research work, it has been built a separate dictionary for the route of administration, symptoms and dosage forms. The words are extracted from the (Vaswani A et.al. 2017). The approach is built by hybrid the dictionary with the deep learning. The detailed architecture of the proposed work has

been presented in the section 4.1. The proposed architecture has several components to classify the entities from the huge amount of literature documents. Section 4.2 presents the hNER (Hybrid – NER) algorithm. Section 4.3 describes the dataset crawling and sentence annotation methods. This section further describes the dictionaries built for identifying the entities. Finally the tools and hardware configuration are described in the section 4.4.

4.1 Architecture Description

In figure 1, the work flow of the proposed hNER algorithm is represented in the architecture format.

The components that are included in the proposed architecture are explained as follows:

4.2 Data Collection

It is the base and important phase of any kind of machine learning work. Identifying and collecting the target data is a challengeable task. To avoid the erroneous data, this work has been targeted on collecting the literature documents from the authenticated web resources. Python based API has been built to crawl the publicly available literature documents from the PubMed based on the certain keywords. Around 200 documents has been downloaded.

4.3 Dictionary

From the literature study, it is clearly evident that, most of the machine learning based model are failed to produce better accuracy due to the lack of dictionaries. To provide a better NER model, this work has been built on the top of new dictionary. The dictionary has developed for the following entities:

- Symptoms
- Route of Administration
- Dosage Forms

These dictionaries are built by collecting the relevant information from multiple resources such as medlineplus , WebMD and fda. The enriched dictionaries are validated by the domain experts.

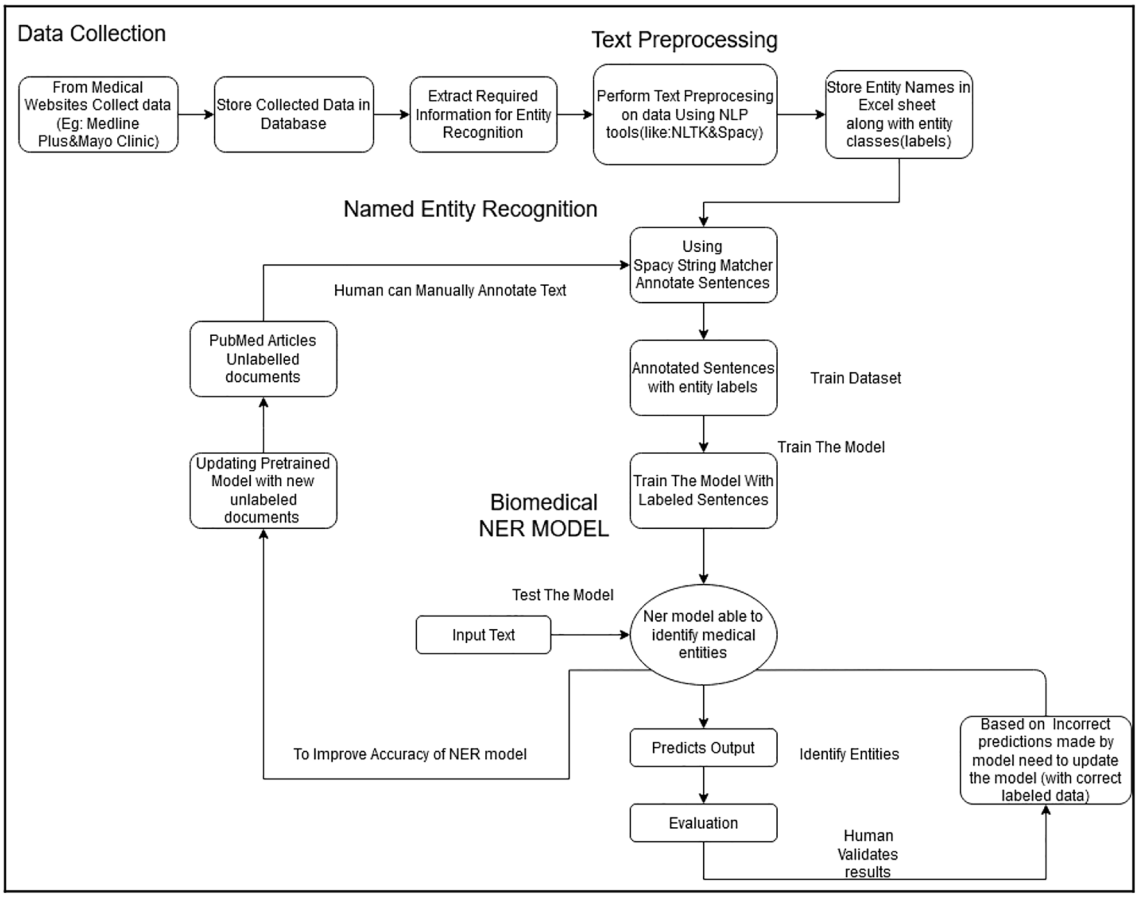

Fig. 1 Flow Diagram of Proposed Hybrid Approach

4.4 Text preprocessing

This is the most important phase of the proposed NER model. Preprocessing is the heart of the proposed architecture which cleaned the raw data and provide the annotated sentence with entities. This step involves in the formation of unstructured data into meaningful format. JSON format has been used for this work. The raw data are tokenized by annotating the sentences with the entity. For training a model annotated sentences are essential. A custom annotated dataset was developed internally for the three entities: Symptoms, Route of Administration and Dosage Forms. The annotated sentences contain the start and end position of the characters. This makes easy to train the model. Figure 2 represents the sample of annotated sentences route and chemical entities. The detailed description of the dataset preparation is described in the section 4.3.

4.5 Named entity recognition

After data preprocessing step, cleaned data is passed in to Bio-NER model. This phase train the model with the annotated sentences. The model is built based on the Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM). The model is retrained by the dataset generated from the dictionary based approach. The SciSpacy model bc5cdr (Mark Neumann et.al. 2019) has been used and retrained by the new annotated dataset. The detailed algorithm of the proposed Hybrid-NER is presented in the Section 4.2.

4.6 Validation

This is the final phase of the proposed architecture. The generated dataset is split into training and testing dataset. 80% of the dataset are being considered as training dataset and 20% are considered as test dataset. The proposed approach is validated by using the dictionary and human interventions. The tool is developed in order to make ease of retraining. The domain expert can identify the correct entity and update the annotated sentences. This will continue until the accuracy is improved. The dictionary based validation is relies on the occurrence of the entity in the dictionaries. The retraining feature in the proposed architecture produced more accuracy than the existing one. The proposed NER approach is evaluated by using the confusion matrix.

4.7 hNER algorithm

In this section algorithm for preparing datasetand building the model is presented. The new dictionaries were built for symptom, dosage forms and route of administration. The extracted sentences are annotated with these dictionaries and divided into test and train dataset. The hybridization of dictionary and retraining the model is named as Hybrid-NER (hNER).

4.7.1 Algorithm: hNER

Input: Preprocessed text documents.
Output: Labelled entities

4.7.2 Method

D^*, set of documents
S^*, set of sentences
A^*, set of annotated sentences
E^*, set of entities
W^*, set of words
GD^*, generated dataset

4.8 Generating dataset

1: **For** d in D:
2: Split document into set of sentences S
3: **For** s in S:
4: **Generate** annotated sentences
5: **Aggregate** all the annotated sentences $\rightarrow GD$
6: **End for**
7: **End for**

4.9 Training and Testing the Model

1: **Load** bc5cdr model
2: **Enable** nlp.resume training
3: **Set** epoch 95

Fig. 2 Annotated Sentences


```
{{
    "content": "for empiric treatment of bacterial sinusitis. Nasal steroids are highly effective for both viral",
    "entities": [ [46, 51,"ROUTE"]]
  }, {
    "content": "ment of LTB is nebulized epinephrine and dexamethasone.",
    "entities": [ [25, 36, "CHEMICAL"], [41, 54, "CHEMICAL"]]
  }}
```

4: **Retrain** the model with the GD
5: **Save** the custom model
6: **Test** the model
7: **Visualize** with the displacy()

The algorithm has been described in two phases as follows: (i) Generating dataset and (ii) Training and testing the model. The generated dataset are stored in JSON format. The stored data is loaded into the python main script for training the model. The proposed approach is been experimented with blank model and existing model. The existing model is retrained using the Transfer Learning concept. CNN play the major role of training the model. The trained model is then saved in a local disk with custom name.

4.10 Dataset preparation

The literature documents around 100 numbers are downloaded by using a python API which has been developed using the python beautiful soup library. The downloaded documents were in PDF format. The retrieved data are converted into raw text. The documents are split into sentences. Spacy phrase matcher is used to annotate the start and end position of the entities. The sentences are further filtered based on the presence of the entities. Table 1 shows the detailed view on the dataset based on the presence of counts of entities. The dataset is fragmented into train-set and test-set created on 80% and 20% respectively. Figure 2 represents the model of annotated data.

4.11 Experiments

The annotated dataset are split as training and testing set randomly into 80 percentage and 20 percentage respectively. The sentences are filtered based on the presence of entities. The empty and duplicate sentence are removed in the pre-processing phase. This reduces the computation time and improve accuracy of model. The annotated sentence are built based on the Goldstandard corpus.

Among the various customized NER model, spacy is one of the powerful resource. It is easy to build a customized NER model. SciSpacy provides bc5cdr NER model to identify the chemical and diseases. This model is pre-trained with 1500 documents. Annotated data is used to retrain the based model which adds more entities to the base model. The entities are added into the existing pipeline of the SciSpacy model, whereas, the model is well categorized with the vocabulary, part-of-speech tagging and NER.

The existing bc5cdr model is trained with the proposed annotated text and saved into disk as new model. Transfer learning which produces better results is discussed in the background study of this paper. Concept of implementing transfer learning reduces the computational cost of building a new model from the scratch. On the other hand, it enables researchers and developers to work with medium configuration. The model is trained using the Convolutional Neural Network (CNN) with Long Short Term Memory (LSTM) methods. CNN-LSTM based training produced better accuracy than the other supervised learning method.

The proposed research work is experimented using the following components to obtain the better accuracy:

- Spacy NLP pipeline on the blank model
- SciSpacy bc5cdr retraining with newly annotated dataset

The experimental setup includes python 3.8 on the 64 bit architecture machine which is operating by Ubuntu 20.04 LTS with 16 GB RAM.

5 Results and Discussions

The proposed approach is experimented on the two models. The first model is newly developed with Spacy blank model used byCNN-LSTM algorithm. Second model is developed by using the transfer learning. The newly annotated data are retrained on the SciSpacy model. The trained model is tested with the test dataset and with real-time dataset. The proposed algorithm is incorporated with the newly developed tool. The tool is designed to estimate the accuracy of the proposed model by two methods as follows: dictionary and human-in-the-loop. The dictionary based method would checked the predicted entities and compare it with the dictionaries. True positive, true negative, false positive, false negative is calculated based on the presence of the objects correctly. Another validation approach is monitored by the

Table 1 Description of Dataset

Dataset	Total Sentences	Entity Counts				
		Chemical	Disease	Symptom	Dosage Forms	Route of Administration
Annotated Sentences	30,774	7157	9391	2580	5412	6234
Train Data	24,620	5726	7513	2064	4330	4987
Test Data	6154	1431	1878	516	1082	1247

domain expert i.e. human-in-the-loop. The domain expert can mark the entities as positive or negative based on their knowledge. It is observed that transfer learning increased the accuracy of the model. The convenience of the training data increase the performance of the NER model. It steadily increase the model performance and reduce the computational time. The results evident that developing blank model would take more time than the time taken for training the pre-trained model. Route of administration and dosage level are the two entities that are newly identified. Pharmacovigilance which majorly working on toxicology reports. These two entities are largely used in generating toxicology report.

The performance of the proposed hybrid approach is evaluated by the familiar measures precision, recall and F-score.

5.1 Precision

It is the division of the entitiesidentified, that are related to the user's information need. The equation for calculating the precision is given in the equation 1.

$$Pr = \frac{tp}{tp + fp} \tag{1}$$

5.2 Recall

It is the fraction of the entities identified that are relevant to the query, which are fruitfully retrieved. The equation for calculating the recall is given in the Eq. (2).

$$Re = \frac{tp}{tp + fn} \tag{2}$$

5.3 F-score

It is a measure that combines precision and recall. It is the harmonic mean of precision and recall, the traditional F-score is given in the Eq. (3).

$$F = 2 * \left(\frac{precision * recall}{precision + recall}\right) \tag{3}$$

5.4 Accuracy

The accuracy of the identified entities is measured by using the following equation. The equation for calculating the accuracy is given in the Eq. (4).

$$A = \frac{tp + tn}{Total} \tag{4}$$

Table 2 Entity-wise Precision and Recall Scores

Entities	Precision		Recall	
	Baseline	hNER	Baseline	hNER
Drugs	64.63	79.10	62.18	77.70
Disease	62.59	76.10	60.14	72.70
Symptom	61.05	73.13	60.87	70.70
Dosage Forms	63.14	74.13	61.70	71.10
Route of Administration	62.43	72.63	60.17	70.14

Table 2 presents the results of precision and recall obtained for the baseline and hNER model. The proposed ML model is trained with 80% of annotated sentences and tested with the 20% of the non-annotated sentences. The evaluation is compared using the actual annotated dataset with the predicted test dataset. Figures 3 and 4 represents the chart diagram of the precision and recall of the experimented models respectively. The chart depicts the increase in the amount of accuracy for the transfer learning method.

Table 3 depicts the F1 scores of the model experimented. The F1 score of the hNER model is higher than the baseline model. It is observed that around 14% is increased for the entities disease and drugs. 11% of increase for the entities symptom, dosage form and route of administration. The bar chart of the Table 3 is shown in the Fig. 5.

Later it was also tested with a PDF document which is not downloaded at the time of dataset preparation. The overall accuracy which calculated by using the Eq. (4) produced 66.4% for all the entities.

6 Conclusion

This research work presented the detailed study of the NER on life science domain. It also highlighted the role of transfer learning to enhance the machine learning model. The proposed hybrid approach identified named entity and outperformed well than the existing baseline method. The transfer learning shows the increase of around 15% accuracy when compared to the baseline method. Baseline model and proposed hNER model was trained with 80% of annotated sentences and tested with 20% of annotated sentence. The F1 score of the experimented model has shown a progressive improvement. As an add-on, the validation tool is more useful to find the accuracy by domain expert. In future, it is plan to boost the quantity of entities. Enriching the dictionaries by adding more object will give more accuracy. The work will be extended to update the dictionary from the suggestions of domain expert and retrain the model instantly.

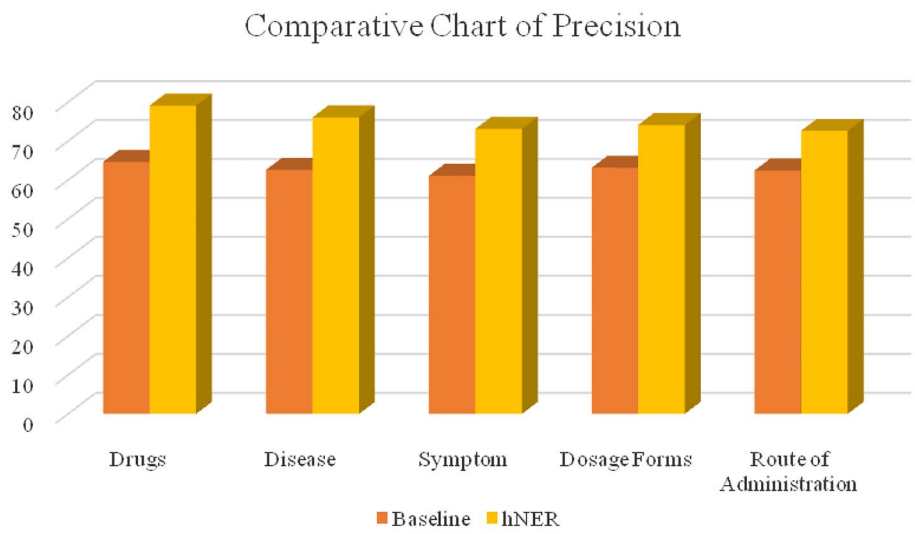

Fig. 3 Comparative Chart of Precision

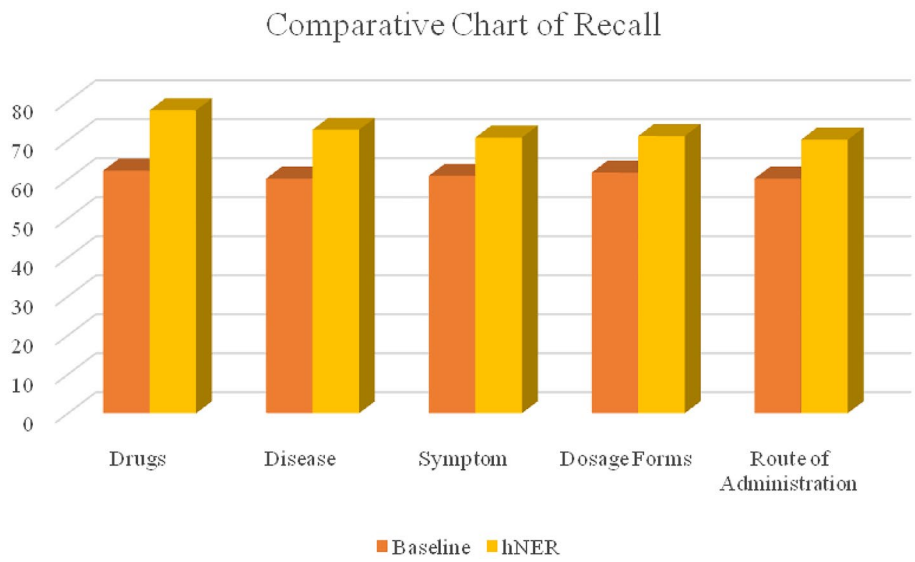

Fig. 4 Comparative Chart of Recall

Table 3 Entity-wise F1 Scores

Entities	Baseline	hNER
Drugs	63.38	78.40
Disease	61.34	74.40
Symptom	60.96	71.89
Dosage Form	62.14	72.89
Route of Administration	61.28	71.36

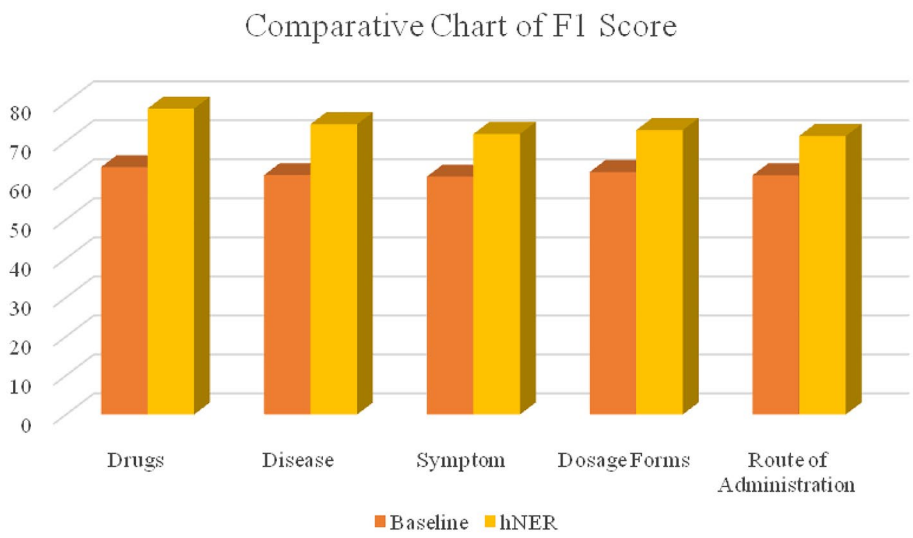

Fig. 5 Comparative Chart of F1 Score

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no Conflict of Interest.

References

Chieu HL, Ng HT. (2002) Named entity recognition: A maximum entropy approach using global information. Pennsylvania: Association for Computational Linguistics. In: Proceedings of the 19th International Conference on Computational Linguistics. 1: 1–7.

Cho H, Choi W, Lee HA (2017) Method for named entity normalization in biomedical articles. Application to diseases and plants. BMC Bioinformatics 18(1):451

Devlin J, Chang MW, Lee K. Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.

Galea D, Laponogov I, Veselkov K (2018) Exploiting and assessing multi-source data for supervised biomedical named entity recognition. Bioinformatics 1:9

Giorgi JM, Bader GD (2018) (2018) Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics 34(23):4087–4094. https://doi.org/10.1093/bioinformatics/bty449

Gridach M (2017) Character-level neural network for biomedical named entity recognition. J Biomed Inform 70:85–91

Grishman R, Sundheim B. (1996) Message understanding : A brief history. In: COLING 1996. The 16th International Conference on Computational Linguistics. Copenhagen; 1996. Volume 1.

Wen G, Chen H, Li H, Yang Hu, Li Y, Wang C (2020) Cross domains adversarial learning for Chinese named entity recognition for online medical consultation. J Biomed Inform. https://doi.org/10.1016/j.jbi.2020.103608

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyeret.(2016) Neuralarchitectures for named entity recognition HLT-NAACL.2016.

Habibi M, Weber L, Neves M, Wiegandt DL, Leser U (2017) (2017) Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 33(14):37–48

Huang Z, Xu W, Yu K. (2015) Bidirectional LSTM–CRF models for sequence tagging arXiv preprint arXiv 1508.01991.

Isozaki H, Kazawa H. (2002) Efficient support vector classifiers for named entity recognition. Pennsylvania: Association for Computational Linguistics. In: Proceedings of the 19th International Conference on Computational Linguistics. 1: 1–7.

Jing Li, Aixin Sun, Jianglei Han, Chenliang Li.(2018) A Surveyon Deep Learning for Named Entity Recognition. CoRR,abs/1812.09449

Kazama JI, Makino T, Ohta Y, Tsujii JI. (2002) Tuning support vector machines for biomedical named entity recognition. Pennsylvania: Association for Computational Linguistics. In: Proceedings of the ACL-02 workshop on Natural Language Processing in the Biomedical Domain 3: 1–8.

Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. (2016) Neural architectures for named entity recognition. arXiv preprint arXiv: 1603.01360.

Leaman R, Islamaj Do˘gan R, Lu Z. (2013) DNorm A disease name normalization with pairwise learning to rank. Bioinformatics 29(22):2909–2917

Leser U, Hakenberg J (2005) (2005) What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform 6(4):357–369

Li Q (2012) Literature Survey: Domain Adaptation Algorithms for Natural Language Processing. Department of Computer Science the Graduate Center, The City University of New York, 2012: 8–10.

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

Mark Neumann,Daniel King, Iz Beltagy, Waleed Ammar (2019) Scispacy: Fast and robust modelsfor biomedical natural language processing .Proceedings of the 18th BioNLP Workshop and Shared Task.

Melamud O, Goldberger J, Dagan I. (2016) Context2vec: Learning generic context embedding with bidirectional LSTM. Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Berlin; 2016:51–61.

Missen MMS, Naeem A, Asmat H (2020) (2020) Improving seller–customer communication process using word embeddings. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-020-02323-

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

Oudah M, Shaalan K.(2012) A pipeline Arabic named entity recognition using a hybrid approach. Proc COLING. :2159–76.

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowledge Data Engineering 22:1345–1359

Rau LF (1991) Extracting company names from text. The Seventh IEEE Conference on Artificial Intelligence Application, Proceedings. Florida: IEEE; 1: 29–32.

Rawassizadeh R, Price BA, Petre M (2015) Wearables: has the age of smart watches finally arrived? ACM Commun 58(1):45–71

Rebholz-Schuhmann D, Yepes AJJ, Van Mulligen EM, Kang N, Kors J (2010) The CALBC Silver Standard Corpus - Harmonizing multiple semantic annotations in a large biomedical corpus. J Bioinform Comput Biol 8:163–179

Rocktaschel T, Weidlich M, Leser U (2012) ChemSpot (2012) A hybrid system for chemical named entity recognition. Bioinformatics 28(12):1633–1640

Schnall A, Heckmann M. 92014) Integrating sequence information in the audio-visual detection of word prominence in a human–machine interaction scenario. In: Fifteenth Annual Conference of the International Speech Communication Association. Singapore; 2014.

Settles B. (2004) Biomedical named entity recognition using conditional random fields and rich feature sets. Barcelona: Association for Computational Linguistics.In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, pp 104–7.

Shah AM, Yan X, Shah SAA (2020) (2020) Mining patient opinion to evaluate the service quality in healthcare: a deep-learning approach. J Ambient Intell Human Comput 11:2925–2942. https://doi.org/10.1007/s12652-019-01434-8

Snow R, O'Connor B, Jurafsky D, Ng AY. (2008) Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In: Proceedings of the conference on Empirical Methods in Natural Language Processing, pp 254–63.

Gajendran S, Manjula D, Sugumaran V (2020) Character level and word level embedding with bidirectional LSTM – Dynamic recurrent neural network for biomedical named entity recognition from literature. J Biomed Inform. https://doi.org/10.1016/j.jbi.2020.103609

Thorne S (2000) (2007) Data analysis in Qualitative Research. Evid-Based Nurs 3(3):68–70

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Polosukhin I (2017) Attention is all you need. Advances in Neural Information Processing Systems, California

Wang X, Yang C, Guan R (2018) A comparative study for biomedical named entity recognition. Int J Mach Learn Cybernet 9(3):373–382

Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big Data. https://doi.org/10.1186/s40537-016-0043-6

Zhao Z, Yang Z, Luo L, Wang L, Zhang Y, Lin H, Wang J (2017) (2017) Disease named entity recognition from biomedical literature using a novel convolutional neural network. BMC Med Genomics 10(5):73

Zhu Q, Li X, Conesa A, Pereira CGRAM-CNN (2017) A deep learning approach with local context for named entity recognition in biomedical text. Bioinformatics 34(9):1547–1554