INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access
RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Named Entity Recognition using Ensemble Learning

R. Ramachandran[1], Dr. K. Arutchelvan[2], R. Senthamizh Selvan[3],
[1,2,3]Assistant Professor, Department of Computer and Information Science, Annamalai University, Tamil Nadu, India.
ramachandranr.au@gmail.com[1]
karutchelvan@yahoo.com[2]
mrsenthamizh@hotmail.com[3]

Abstract

Upgrading Industry 4.0 to 5.0 provides numerous research opportunities for the industrialists and researchers. This industrial revolution cross the peak of automation in the life science domain. In this digitalized world, big data plays a key role to provide the valuable insights by using various analytical methods. In life science, available of huge textual data contains wide spread of valuable information. To extract the hidden information from the big data, natural language processing plays a major and significant role. In NLP, named entity recognition is one of the key factor and biggest challenge for the research community. This paper presents the high level architecture of NER using ensemble learning method. The EL model contains a dictionary based entity identifier and a self-learning classifier. Proposed model outperformed well and produced high accuracy.

Keywords: Natural Language Processing, Named Entity Recognition, Conditional Random Field, Lexicon Based Approach, Ensemble Learning

1.Introduction

Life science is one of the prominent and growing domain in the business industry. The revolution of Artificial Intelligence (AI) in the pharmacy industry provided many research and job opportunity. Plenty of applications that are being used in life science industry are migrated to automation. Most of data are in textual format and creates the biggest challenge to the researchers and industrialist. To work with the textual data, NLP is one of the key technique to extract the valuable insights.

In various industries such as healthcare, education, finance, social media, etc., contains abundant information which are difficult to handle. NLP is significant to handle those sources. This paper distressthe role of NLP strategies in biomedical field.According to statista , in 2019 they have projected that in 2020 the healthcare industry has 2134 Exabyte of data.The data deluge in healthcare industry which is commonly generated by electronic healthcare record are stored inregional language. The stored data are in organized structure which is making more difficulties for the retrieving the hidden information from the huge amount of text data.

The digitalized information of the clinical records are frequently store in the formal language. NLP is helpful for researchers and industrialists to communicate occasions and clinical ideas, astonishingly it makes the information hard for looking due to the lack of technologies and tools.

To overcome these difficulties, the data must be properly processed by the NLP techniques. Named Entity Recognition (NER) is a key NLP errand to extricate the elements of intrigue (e.g., ailment names, medicine names and lab tests) from clinical

stories, along these lines to help clinical and translational exploration.

The paper is organized as follows: the background study of the NER in biomedical domain is present in the section 2.The proposed architectural flow the Lex-NER model is described in the section 3 with elegant workflow figures. Section 4 presents the results and discussions. Section 5 concludes with limitations of the proposed work

2. Background Study

Named Entity Recognition (NER) is a powerful technique in the NLP [1]. It is a sub-field of information retrieval.It is an errand of perceiving the articulations that ought to be ordered as articulations indicating substances. Model substance tags in clinical arena are ailments, drugs, treatment, qualities, malignant growth, protein and RNA [2, 3, and 4]. A great part of the examination in life science informatics has focused on NER. As indicated by [5] the majority of the techniques are rule-based, in spite of the fact that there are executed some half and half methodology that consolidate AI with these principles.

The creators in [6] makes reference to Conditional Random Fields (CRF), Support Vector Machines (SVM) and Hidden Markov Model (HMM) as regular AI strategies that are at present applied for NER undertakings in clinical space. The latest papers focus on profound wisdom methodologies put on repetitive neural systems (RNNs), for example, Long-Short Term Memory (LSTM) [7], Gated Recurrent Units (GRU) [8]. Basic pattern is joining the RNN with factual technique on head of the intermittent layers. It guarantees that the ideal succession of labels over the whole sentence is acquired [9]. CRF is the most regularly utilized measurable strategy in this cross breed approach. The creators of [9] consolidated RNN with CRF. Because of the difficulties recorded beneath the Clinical NER endeavors get lower execution estimates esteemed thebest F1 score acquired by [9] is sums 91.32% in correlation of comparable preliminaries with corpuses in non-specialized fields, where as of late the creators of [10] got F1 score ninety three percentile on the CoNLL 2003 corpus.

Right off the bat, the information accessible for scientists in the biomedical field is restricted, for the most part because of the patient security and classification necessities. The accessible clarified databases are normally inadequate for named element acknowledgment undertaking to prepare the model [6]. Also, the clinical writings are written in a particular way, unique in relation to customary language. There are a great deal of inadequate sentences, casual syntax and covered with incorrect spellings and non-standard shorthand, shortened forms and abbreviations. Also, the medication is a quickly extending field with huge number of investigates led the add to continually developing number of clinical ideas. It makes incredibly hard to staying up with the latest. Besides, ideas in medication regularly convey importance, identified with the idea. It infers the NER models to keep the word setting data along the preparation procedure. Another ordinary element is that clinical language is described by long expressions containing exceptional characters and runs.

The greater parts of the investigations are performed on Corpuses in English. Second most well-known language is Chinese. There are obviously inadequate with regards to explores in different dialects. The term characteristic language is utilized to depict any language utilized by people, to recognize it from programming dialects and information portrayal dialects utilized by PCs and portrayed as counterfeit [11, 12]. Normal language handling (NLP) term depicts computational methods that procedure communicated in and composed human language [13]. Characteristic language handling incorporates information preprocessing strategies like information cleaning, tokenization, standardization (stemming, lemmatization or different types of normalization). Setting up the content requires picking the ideal apparatuses; anyway it assists with improving precision of continuing NLP errands. Different assignments of NLP focus on removing the factual highlights like term frequency, inverse document frequency or linguistic highlights including Part of Speech (POS) labeling. NLP methods are devices to accomplish the unrivaled errand. Data Extraction (IE) including scanning for pertinent data in records exist among the most applied assignments.

NER is a phase of Information Extraction. It is one of key NLP undertakings that assists with changing over unstructured content into PC coherent organized information [13]. NER alludes to the undertaking of perceiving the articulations

indicating substances (for example Named Entities, for example, illnesses, medications or individuals' names, in free content archives [14]. NER can be tackled with the utilization of numerous methods that can be separated into a few gatherings [15]: word reference based methodology, rule based methodology, measurable methodology, profound learning approach, crossover approach. The author performed NER task on GENIA corpus. Genia is normally utilized corpus by analysts both as word reference and as base corpus to perform NER task. The NER model is accessible in various variants and various configurations. Authors have utilized corpus which comprises of 1001 dynamic records from MEDLINE database and it is a scientific classification of 30 organically pertinent classes.

3. Materials and Methods

Lex-NER is the hybrid NER framework which istrained and tested on PubMedabstracts. The proposed work aims to build a hybrid model which combines both the string matcher and Machine Learning (ML) model to produce better accuracy. The ML which is incorporated a phase known as human-in-the-loop. The human-in-the-loop phase is used to increase the accuracy of the ML model. The domain experts evaluate the results of ML model and update the training dataset. This helps to increase the accuracy level.

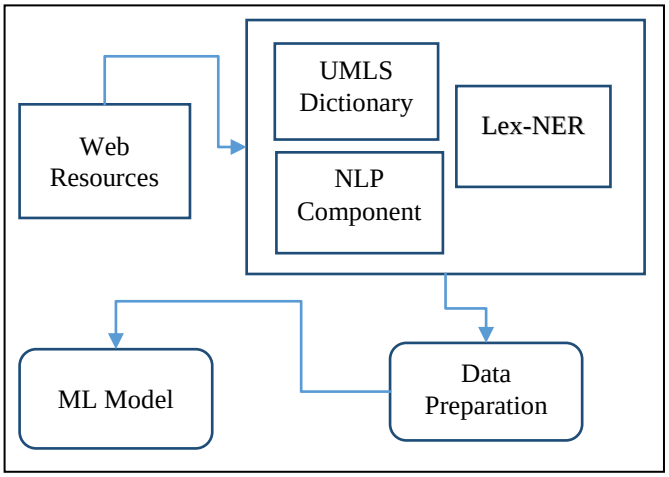

Fig.1.High-level Architecture of Modules in NER Processing

Figure 1 presents the high level architecture of the Modules in NER processing. The proposed framework is composed of the following four modules.

(1) Identifying the abstracts based on the keywords
(2) NLP module which includes the text processing
(3) Data preparation for training the machine learning model
(4) Machine Learning module

The first module is to identify the abstracts from the PubMed articles. For this experiment, it has been limited with certain keywords. Keywords such as drug names, diseases and symptoms. Based on the keywords the document is classified. Around 531 documents are collected for the proposed work.

Second module, is the engine of the proposed work. It contains the Lex-NER component, dictionary component and NLP component. The dictionary component is built by the UMLS dictionary. UMLS is one of the powerful database which contains around 130 attributes of different tokens. The texts in the documents are processed by using the basic NLP component. Figure 2 presents the workflow of Lex-NER component. The common techniques of NLP component are given in the figure 2. The important keyword is extracted and matched by using the string matcher. The matched word from the UMLS dictionary is then passed into Lex-NER component.

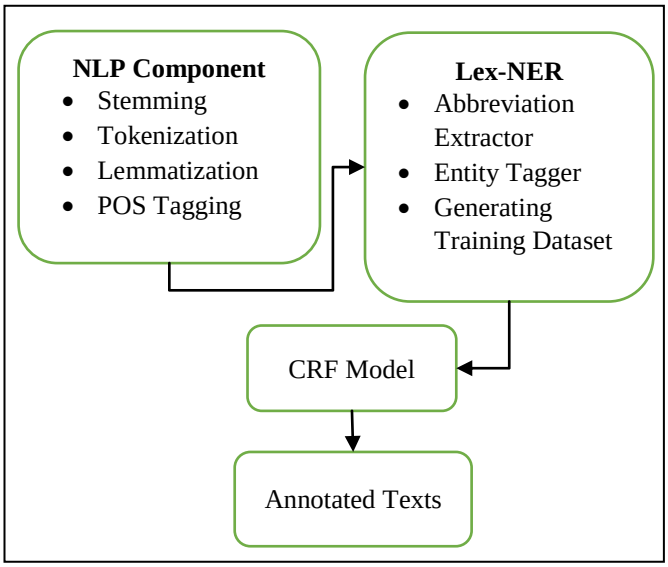

Fig.2. Workflow of Lex-NER

Third module is data preparation for training the ML model. The training data set is generated and stored in JSON format. It the light-weight format

to train the model with huge datasets. The following is an example of training datasets.

[('caineiontophoresis system for topical anesthesia in adults and children: a randomized, \n', {'entities': [(20, 26, 'DRUG'), (31, 38, 'ROUTE')]}), ('caineiontophoresis system for topical anesthesia in adults and children: a randomized, \n', {'entities': [(20, 26, 'DRUG'), (31, 38, 'ROUTE')]}), ('Kreyden, O.P. Iontophoresis for palmoplantar hyperhidrosis. \n', {'entities': [(45, 58, 'DISEASE')]}), (' Dermal, subdermal, and systemic concentrations of granisetron \n', {'entities': [(51, 62, 'DRUG')]}), ('Morrel, E.M., Spruance, S.L. & Goldberg, D.I. Topical iontophoretic administration of \n', {'entities': [(46, 53, 'ROUTE')]}), ('acyclovir for the episodic treatment of herpes labialis: a randomized, double-blind, \n', {'entities': [(0, 9, 'DRUG')]})]

In the proposed work, the CRM (Conditional Random Field) based ML model is used to train the annotated texts.The following equation 1 states the formula of the CRF model. Y is the hidden state and X is the entities which are observed by the string matcher. y_t, y_{t-1}, X_tdenotes the features of the datasets and $\theta_k f_k$resembles the weight of the features. The weight of the feature is calculated by the maximum likelihood estimation. The feature are set by the developer.

$$p(y|x) = \left(\frac{1}{Z(x)}\right) \prod_{t=1}^{T} exp\{\sum_{k=1}^{K} \theta_k f_k(y_t, y_{t-1}, X_t)\} \quad \ldots. (1)$$

Table 1 depicts the results and discussion of the proposed work.

Table.1. Comparative Results of Lex-NER

Comparative study			
Techniques	**Precision**	**Recall**	**F-Score**
LingPipe	72.95	88.49	79.97
ABNER	86.93	51.49	64.88
Lex-NER	88.66	84.32	86.43

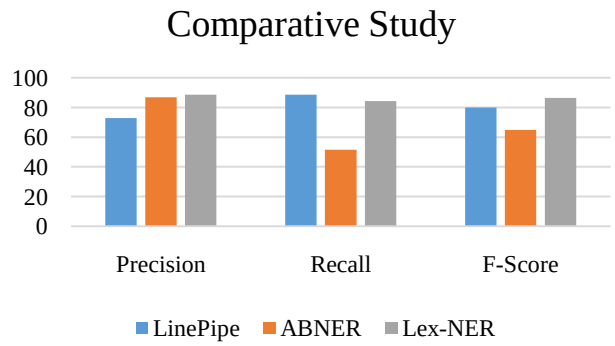

Fig.3. Comparative Study of NER methods

The comparative study of proposed Lex-NER model with LinePipe and ABNER is presented in the table 1. From the table it is evident that Lex-NER model outperforms well than the existing models. Figure 3 depicts the graphical representation of the comparative study.

Conclusions

The proposed hybrid based NER model outperforms well when compared to the existing LinePipe and ABNER. The state of art which has examined and presented in the background study provide the solid knowledge for the NER. The model comprises of the string matcher and CRF model. The UMLS database is used to identify the entities with the phrase matcher. The matched words are annotated by the entity and passed into model. The CRF model which has been worked based on forward parsing method produced good accuracy. The model is trained using on the 531 abstracts which has been extracted from the PubMed database. In future, the work would be extended on increasing the abstracts and increasing number of features.

References

[1] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler and J. F. Hurdle, "Extracting information from textual documents in the electronic health record: A review of recent research", pp. 128–144, ISSN 0943-4747.

[2] A. B. Abacha and P. Zweigenbaum, "Medical Entity Recognition: A Comparison of Semantic and Statistical Methods", Proceedings of BioNLP 2011 Workshop, BioNLP'11, pp. 56–64, 2011.

[3] V. Hatzivassiloglou, P. A. DubouÃl' and A. Rzhetsky, "Disambiguating proteins, genes,

and RNA in text: A machine learning approach" Suppl 1:S97–106. ISSN 1367-4803.

[4] J. Song, B. Jo, C.Y. Park, J.-D. Kim and Y.-S. Kim, “Comparison of named entity recognition methodologies in biomedical documents”, BioMedical Engineering OnLine, Vol. 17(2):158, 2018.

[5] S. Pradhan, N. Elhadad, B. R. South, D. Martinez, L. Christensen, A. Vogel, H. Suominen, W. W. Chapman and G. Savova, “Evaluating the state of the art in disorder recognition and normalization of the clinical narrative”, Vol. 22(1), pp. 143–154, ISSN 1527-974X.

[6] J. Zhang and J. Li and S.Wang and Y. Zhang and Y. Cao and L. Hou and X. Li, "Category Multi-representation: A Unified Solution for Named Entity Recognition in Clinical Texts", Advances in Knowledge Discovery and Data Mining, Springer, pp. 275–287, 2018.

[7] Y. Qin and Y. Zeng, "Research of Clinical Named Entity Recognition Based on Bi-LSTM-CRF", Journal of Shanghai Jiaotong University (Science), Vol. 23(3), pp. 392–397, 2018.

[8] A. P. Quimbaya and A. S. MÃžnera and R. A. GonzÃ¸alez Rivera and J. C. DazaRodrÃ_guez and O. M. MuÃ´sozVelandia and A. A. Garcia PeÃ´sa and C. LabbÃl’, "Named Entity Recognition Over Electronic Health Records Through a Combined Dictionary-based Approach", Procedia Computer Science, Vol. 100, pp. 55–61, 2016.

[9] J. Qiu and Q. Wang and Y. Zhou and T. Ruan and J. Gao, "Fast and Accurate Recognition of Chinese Clinical Named Entities with Residual Dilated Convolutions", Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 935–942, 2018.

[10] A. Baevski and S. Edunov and Y. Liu and L. Zettlemoyer and M. Auli, "Cloze-driven Pretraining of Self-attention Networks", http://arxiv.org/abs/1903.07785.

[11] G. Lample and M. Ballesteros and S. Subramanian and K. Kawakami and C. Dyer, "Neural Architectures for Named Entity Recognition, http://arxiv.org/abs/1603.01360.

[12] D. Jurafsky and J. H. Martin, "Speech and Language Processing", 2nd Edition. Prentice Hall, ISBN 978-0-13-187321-6.

[13] Y. Sasaki and Y. Tsuruoka and J. McNaught and S. Ananiadou, "How to make the most of NE dictionaries in statistical NER", Vol. 9(11):S5, ISSN 1471-2105.

[14] M. A. Hearst, "Untangling Text Data Mining", Proceedings of 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, ACL’99, pp. 3–10, 1999.

[15] M. Allahyari and S. Pouriyeh and M. Assefi and S. Safaei and E. Trippe and J. B. Gutierrez and K. Kochut, "A Brief Survey of Text Mining: Classifiation, Clustering and Extraction Techniques, 2017.