

NER-DD: A Named Entity Recognition Tool for Tagging Drugs in Disease-related Documents

R. E. Ramos-Vargas*, J. Monroy-Vargas,
I. Román-Godínez¹, S. Torres-Ramos¹,

¹Departamento de Ciencias Computacionales, Universidad de Guadalajara, Jalisco, México

**rigo.ramos@alumnos.udg.mx*

Abstract

The amount of biomedical documents is increasing daily. Therefore, several tools intent to identify automatically important information in those documents such as genes, disease, drugs among others. Depending on the tool selected, the combination of the type of input documents, tagging model, and output information presents some disadvantages. This work attempt to gather three functionalities, the capability of use not only PubMed abstracts but also PDF or TXT files, a neural-based named entity recognition approach, and display co-occurrence tables. Therefore, we present NER-DD a tool for tagging drugs in disease-related documents to determine possible associations between them.

Keywords: BioNER, drug-disease association, drugs tagger, web tool.

1. Introduction

PubMed is one of the most used available databases about biomedical literature [1], and according to statistical reports, by January 2020 there were more than 30 million records on their platform [2]. This huge quantity of data is evidence of the importance and increasing interest in the biomedical research area.

As a result, it is being many efforts directed not only developing information retrieval systems but also to identify biomedical entities on documents automatically. Some of those efforts result in several web tools such as PubTerm [3], PubTator [4], and ezTag [5], which main intention is to tag biomedical entities such as diseases, drugs, genes, among others, in PubMed documents.

Particularly, PubTerm and PubTator can tag abstracts or complete articles obtained from PubMed or PubMed central but they are not able of doing that on an alternative source. On the other hand, even though ezTag can tag documents from different sources, this tool required the user to format the entry through a particular tool, requiring specialized skills.

In addition, according to [6], these latter tools present some limitations regarding the use of older Named Entity Recognition (NER) models resulting in a limited identification of newer biomedical entities. Therefore, in [6] the tool BERN was presented, an artificial neural network approach that combines BiLSTM-CRF and multi-type normalization to address such a problem.

However, BERN does not allow establishing some kind of relation/association between two biomedical entities, which is an important strategy in several tasks such as Drug Repositioning [7][8]. This task has the objective to identify new indication from existing drugs or the application of newly developed drugs in different treatments for they were designed, and for which it is not only important to tag diseases or drugs in documents but also to determine if exists some co-occurrences among those entities.

In this work, we present NER-DD, a web tool for tagging drugs in disease-related documents using

an artificial neural network-based model. NER-DD is a scalable, usable, and friendly web tool that allows users to request abstracts from PubMed and upload documents related to diseases, to compute a co-occurrence matrix that could help in the identification of possible drugs-disease associations.

2. Materials and methods

NER-DD is designed using a modular architecture in order to make it scalable and easy to maintain. It was developed with the programming language Python for some backend functionalities and Django along with Bootstrap for the web domain logic and frontend capabilities.

Fig. 1 shows the four modules of the NER-DD system, which are presented as follows.

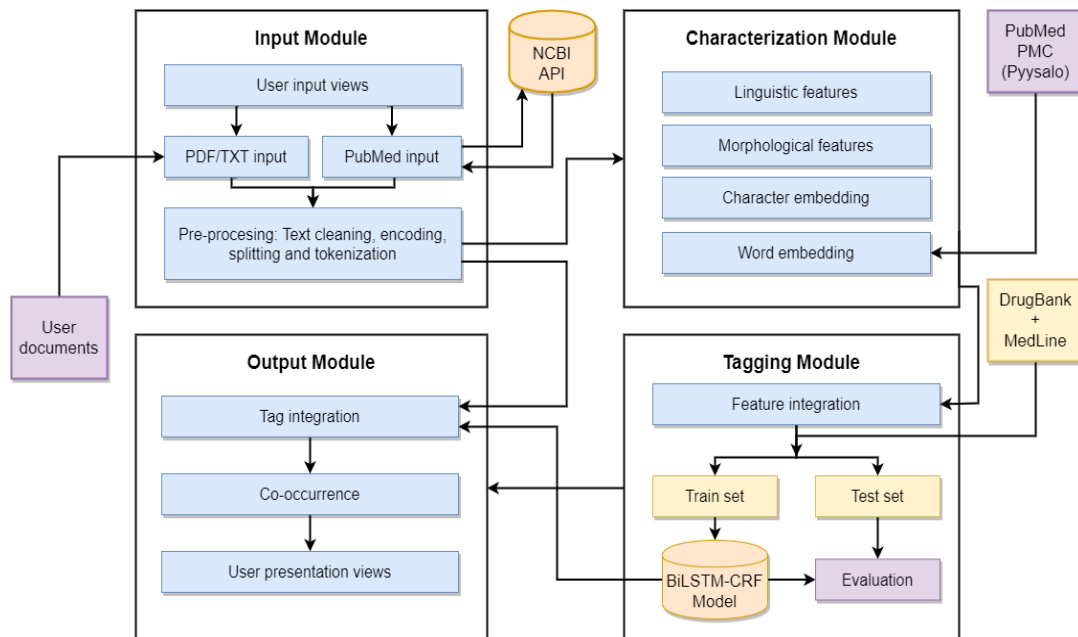


Fig. 1. Module diagram of NER-DD system.

2.1 Input module

The main entrance of the system is the PubMed search view (see Fig. 2), this view allows the users specify the diseases to be searched, either select from the list of the ICD-10 [9], or can write their search terms. The users also provide the number of abstracts per disease to be fetched, and optionally, the date range.

The screenshot shows the PubMed search interface with the following elements:

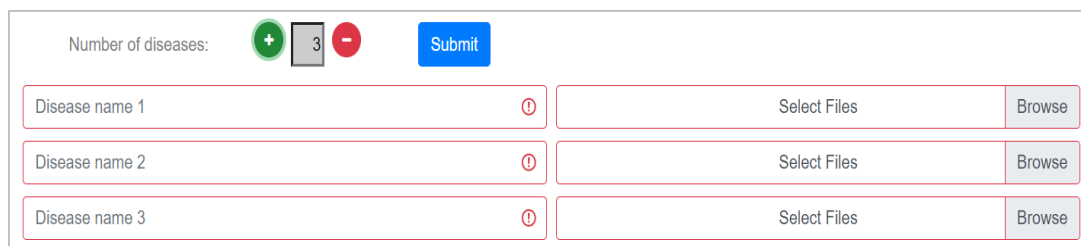
- Diseases list:** A text input field with the placeholder 'Write the disease and press ENTER or choose the option (Max 10)'.
- Number of Abstracts:** A numeric input field containing the value '1' with a green checkmark on the right.
- Start date:** A date input field with the format 'mm/dd/yyyy', a calendar icon, and a green checkmark.
- End date:** A date input field with the format 'mm/dd/yyyy', a calendar icon, a green checkmark, and an information icon.
- Search:** A prominent blue button.
- Enable Dates:** A green toggle switch that is currently turned on.

Fig. 2. PubMed search view.

Once all search parameters have been set, a query to the PubMed database is performed using the NCBI API [10]. The system will be able to retrieve a maximum of 250 articles per disease because this is the maximum quantity allowed from NCBI.

As a result, the query returns a JSON file with the articles IDs that match the search parameters. In case of the date range is not specified, the most recent articles IDs are retrieved. Those IDs are used to request an XML file with all the articles' information (including the abstract).

Additional to the PubMed input, the user can upload PDF or TXT files to be processed. In this case, the system allows associating a set of documents with a particular disease (at most 10 diseases), as shown in Fig. 3.



The screenshot shows a web interface for managing documents. At the top, there is a control for the number of diseases, with a green plus button, a grey box containing the number '3', and a red minus button. To the right is a blue 'Submit' button. Below this is a table with three rows. Each row has a text input field for a disease name (labeled 'Disease name 1', 'Disease name 2', and 'Disease name 3'), a red circular icon with a white 'i' to its right, a 'Select Files' button, and a 'Browse' button.

Fig. 3. User documents view.

The XML, TXT, or PDF documents received are pre-processed to provide the correctly formatted data for the entire system. For each XML file, the text corresponding to the abstract (target text) is extracted using regular expressions, along with the ID, title, authors, release date, and DOI. On the other hand, the plain text from TXT and PDF files (also target text) is extracted and encoded to UTF-8. The PyPDF2 library [11] was used to read the PDF files. Finally, the target text is split into sentences and tokenized into words through NLTK [12] to provide the input data for the characterization and tagging modules.

2.2 Characterization module

The characterization module is used to obtain the representative feature set of the input data. The feature set consists of linguistic and morphological features, character embedding, and word embedding. These features have been used in recent research [13][14] to characterize biomedical entities.

Two linguistic features were used, the token itself and Parts of Speech (POS). Through POS tagging is possible to associate each token with a particular grammatical category based on its context. To extract morphological features of the tokens eight hand-crafted rules were performed: a) is all lower case, b) has first letter capitalized, c) ends with 's', d) contains digits, e) is numeric, f) is alphabetic, g) is alphanumeric, and h) is a stop word. If a given token fulfills a particular rule, then its value is set to one (1), zero (0) otherwise.

The character embedding approach has been used to help on the recognition of biomedical entities by identifying prefixes and suffixes [15]; for example, in the name of a particular drug, the words -zosin and -cycline represent suffixes. At the beginning of the character embedding process, each forward and backward word substring is initialized by a random vector, then, the embeddings are passed one-by-one to a BiLSTM (see Tagging module for a BiLSTM short explanation), then, after several iterations of the BiLSTM, the resulting is the encodings for the beginning and end of each word.

The word embedding implementation used in this work is PubMed-PMC [16], the latter due to it has shown improvements in the performance on the NER task for specific biological domains [17].

Word embedding transforms words into numerical vectors that capture semantic and syntactic regularities [18] expecting that semantically similar words have similar vectors.

2.3 Tagging module

This module is the core of the system. It was implemented through the machine learning technique called NER. The selected algorithm was a Bidirectional Long Short Term Memory (BiLSTM) with a Conditional Random Field (CRF) layer: BiLSTM-CRF.

The BiLSTM is a kind of Recurrent Neural Network (RNN) capable of learning long-term dependencies from sequential data [19]. The CRF layer of the BiLSTM-CRF performs a Viterbi joint decoding of the input sequence [15], achieving a better performance in comparison with the BiLSTM [14]. This neural network algorithm has the best F-Scores in recent published works [13][14] for biomedical named entity recognition.

To train the model, the parameter configuration was set with a Dropout of 0.5, a learning rate 0.01, using the stochastic gradient descent as the optimization algorithm, and a hidden layer dimension of 100 and 25 for Word BiLSTM and Char BiLSTM models, respectively.

The corpus used for training and test the tagging module was DDI [20]. This corpus was used as the gold standard in the SemEval-2013 DDI Extraction Task, particularly in the 9.1 challenge for recognition and classification of pharmacological substances [21], and has been widely used for the drug named-entity recognition task [13][14][22]. The DDI corpus results from the concatenation of 1025 documents from two different sources: DrugBank database and MedLine. The training and testing sets were split as recommended in SemEval-2013 [21].

To perform the training and testing of the tagger module, each token in the training corpus was labeled according to the I-O-B tagging scheme.

2.4 Output module

The output views were built to provide the user with interesting information such as abstract retrieved, labeled drugs, and drugs-disease co-occurrence. Details about each one are presented in the Results section. One of the functionality provided by the output module is the download of the results in several formats, such as, JSON, xml, csv, txt, sql, and excel.

3. Results

NER-DD system is available on this [site](#). The tagging module was tested according to the characteristics described in section 2.3, obtaining 82.97%, 85.31%, and 84.12% for precision, recall, and F-score, respectively. In addition, the following results views are presented: abstracts information, drugs labeled, and drugs-disease co-occurrence. In order to describe each one, a query was performed using the terms *Influenza*, *Dengue*, and *Coronavirus*, retrieving 30 abstracts per disease between 2020-07-06 and 2020-07-16.

Fig. 4 depicts the view corresponding to the retrieved abstracts showing a table per disease. In this particular example, the table shows the last three results retrieved for *Coronavirus*. The table shows information such as "Id Article" corresponding to the PubMed identifier; the "Title" of the article, which is a link to the full article in PubMed repository; and under the title, its author list. The column "Drugs Found" presents the number of drugs identified in each abstract, and in the last column, the "Go to info" button allows access to the tagged abstract, detailed below.

Dengue (1 of 3)			
Influenza (2 of 3)			
Coronavirus (3 of 3)			
Search <input type="text"/> ⌵ ⬇			
Id Article	Title	Drugs Found	Go to info
32667665	Transcriptome-based drug repositioning for coronavirus disease 2019 (COVID-19). Jia Z., Song X., Shi J., Wang W., He K.,	4	Go to info
32667604	The coronavirus and the História, Ciências, Saúde - Manguinhos blog. Cueto M.,	0	Go to info
32667579	Nursing appeals on social media in times of coronavirus. Forte ECN., Pires DEP.,	0	Go to info

Showing 28 to 30 of 30 rows 3 rows per page < 1 ... 6 7 8 9 10 >

Fig. 4. Retrieved abstracts view.

Fig. 5 shows the tagging view with a title, abstract, authors, and publication date, where the highlighted green words of the abstract depict the identified drugs. To control the information displayed, the user has a combo box to choose the disease to be observed and the buttons "Previous" and "Next" to move over abstracts.

Previous		Coronavirus ⌵	Next
Title	Transcriptome-based drug repositioning for coronavirus disease 2019 (COVID-19).		28 of 30
Abstract	<p>The outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) around the world has led to a pandemic with high morbidity and mortality. However, there are no effective drugs to prevent and treat the disease. Transcriptome-based drug repositioning, identifying new indications for old drugs, is a powerful tool for drug development. Using bronchoalveolar lavage fluid transcriptome data of COVID-19 patients, we found that the endocytosis and lysosome pathways are highly involved in the disease and that the regulation of genes involved in neutrophil degranulation was disrupted, suggesting an intense battle between SARS-CoV-2 and humans. Furthermore, we implemented a coexpression drug repositioning analysis, cogenia, and identified two antiviral drugs (saquinavir and ribavirin) and several other candidate drugs (such as dinoprost, dipivefrine, dexamethasone and (-)-isoprenaline). Notably, the two antiviral drugs have also previously been identified using molecular docking methods, and ribavirin is a recommended drug in the diagnosis and treatment protocol for COVID pneumonia (trial version 5-7) published by the National Health Commission of the P.R. of China. Our study demonstrates the value of the cogenia-based drug repositioning method for emerging infectious diseases, improves our understanding of SARS-CoV-2-induced disease, and provides potential drugs for the prevention and treatment of COVID-19 pneumonia.</p>		
Authors	Jia Z., Song X., Shi J., Wang W., He K.,		
Date	2020/Jun/01		

Fig. 5. Tagging view.

To analyze the data, we present two different co-occurrence views. First, Fig. 6 depicts a sample of the resulting table corresponding to the times that a particular drug occurs in a group of documents related to a specific disease. For example, the row corresponding to *Ribavirin* indicates that this drug was found two times in Coronavirus-related abstracts, zero times in Influenza-related abstracts, and zero times in Denge-related abstracts; with a total of 2 occurrences.

	Coronavirus	Influenza	Dengue	Total of occurrences
Heparin	0	1	0	1
interferon	0	0	1	1
Chloroquine	0	1	0	1
Ribavirin	2	0	0	2
Dexamethasone	1	0	0	1
Metformin	2	0	0	2
Methylprednisolone	0	0	1	1
Telemedicine	1	0	0	1
Saquinavir	1	0	0	1
Nonanoic acid	0	1	0	1
Valproic acid	0	1	0	1
Insulin	3	0	0	3
Azithromycin	0	1	0	1
Hydroxychloroquine	3	1	0	4

Fig. 6. Drug occurrences per disease.

Second, Fig. 7 shows a sample of the number of abstracts related to a disease where a drug has been found. For example, for *Coronavirus* disease, eight drugs were found, from which *hydroxychloroquine* happen in three different abstracts while the others drugs occurred in one.

Diseases	Drugs	Number of Abstracts
Coronavirus	telemedicine	1
	saquinavir	1
	dipivefrine	1
	insulin	1
	hydroxychloroquine	3
	ribavirin	1
	metformin	1
	dexamethasone	1
Influenza	nonanoic acid	1
	valproic acid	1
	azithromycin	1
	hydroxychloroquine	1
	fine	1
Dengue	interferon	1
	methylprednisolone	1
	dengue	1

Fig. 7. Number of abstracts where a drug occurs per disease.

The main difference between NER-DD and other state-of-art tools resides in two functionalities that would be interesting for users: 1) Co-occurrence tables that allow the visualization of possible disease-drugs associations, and 2) PDF files tagging that gives the user the possibility to upload full papers in PDF format obtained from different sources.

On one hand, and concerning to the tagging method, BERN is the most similar tool to our proposal, implementing a neural-based NER model, reporting an F-Score of 91.41%. However, BERN uses an additional normalization schema to improve its results and does not present co-occurrence tables or upload pdf files.

On the other hand, ezTag provides access to the PubMed abstract and the capability of upload files. However, the system requires that those files should be preprocessed using a specific programming library, thus, needing users to have some degree of computational skills. Besides, ezTag does not present co-occurrence matrices. Finally, PubTator and PubTerm only perform tagging on PubMed documents fetched using NCBI API.

4. Conclusions

NER-DD is a named entity recognition web tool for tagging drugs in disease-related documents that can be useful in biomedical research to determine possible drugs-disease associations. NER-DD is friendly, scalable and gathers different functionalities of several tools available on the state-of-art; first, a neural-based named entity recognition approach; second, a visual interface that shows drugs-disease co-occurrence tables, without the need for manually doing it; and third, the capability of use PDF or TXT files. At last, even though NER-DD is originally conceived to relate drugs with diseases, the search terms could be anything else, so, the co-occurrence tables could help to make evident drug associations with any other concepts. As future work, the system can be upgraded to recognize several biomedical entities such as genes, diseases, among others. In addition, to improve the tagger performance, it could be implemented a user annotation module and combine several methodologies for named entity recognition.

Conflict of interest declaration

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgements

The author R. E. Ramos-Vargas express his gratitude to CONACyT for the scholarship to pursue his postgraduate studies.

References

- [1] Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature", *Database*, 2011.
- [2] National Institutes of Health, "Statistical Reports on MEDLINE®/PubMed® Baseline Data". [online]. 2020. Available at: <https://www.nlm.nih.gov/bsd/licensee/baselinestats.html>

- [3] J. Garcia-Pelaez, D. Rodriguez, R. Medina-Molina, G. Garcia-Rivas, C. Jerjes-Sánchez, V. Trevino, “PubTerm: a web tool for organizing, annotating and curating genes, diseases, molecules and other concepts from PubMed records”, *Database*, 2019.
- [4] C. Wei, H. Y. Kao, Z. Lu, “PubTator: a web-based text mining tool for assisting biocuration”, *Nucleic Acids Res.*, vol. 41, no. W1, pp. W518-W522, 2013.
- [5] D. Kwon, S. Kim, C. H. Wei, R. Leaman, Z. Lu, “ezTag: tagging biomedical concepts via interactive learning”, *Nucleic Acids Res.*, vol. 41, no. W1, pp. W523-W529, 2018.
- [6] D. Kim, J. Lee, C. So, H. Jeon, M. Jeon, Y. Choi, J. Kang, “A neural named entity recognition and multi-type normalization tool for biomedical text mining”, *IEEE Access*, vol. 7, pp. 73729-73740, 2019.
- [7] J. S. Shim, J. O. Liu, “Recent advances in drug repositioning for the discovery of new anticancer drugs”, *Int. J. Biol. Sci.*, vol. 10, no. 7, pp. 654, 2014.
- [8] H. Xue, J. Li, H. Xie, Y. Wang, “Review of drug repositioning approaches and resources”, *Int. J. Biol. Sci.*, vol. 14, no. 10, pp. 1232, 2018.
- [9] World Health Organization, “International Statistical Classification of Diseases and Related Health Problems, 10th revision. [online]. 2016. Available at: <https://icd.who.int/browse10/2016/en>
- [10] National Center for Biotechnology Information, “Entrez Programming Utilities Help”. [online]. 2010. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK25501/>
- [11] M. Fenniak, “PyPDF2”. [online]. 2014. Available at: <https://github.com/mstamy2/PyPDF2/>
- [12] E. Loper, S. Bird, “NLTK: the natural language toolkit”, in *Proc. ACL*, 2002.
- [13] E. Batbaatar, K. H. Ryu, “Ontology-based healthcare named entity recognition from Twitter messages using a recurrent neural network approach”, *Int. J. Environ. Res. Public Health*, vol. 16, no. 19, p. 3628, 2019.
- [14] I. J. Unanue, E. Z. Borzeshi, M. Piccardi, “Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition”, *J. Biomed. Inform.*, vol. 76, pp. 102-109, 2017.
- [15] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, “Neural architectures for named entity recognition”, *arXiv preprint arXiv:1603.01360*, 2016.
- [16] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, S. Ananiadou, “Distributional semantics resources for biomedical text processing”, In *Proc. LBM*, pp. 39-44, 2013.
- [17] M. Habibi, L. Weber, M. Neves, D. L. Wiegandt, U. Leser, “Deep learning with word embeddings improves biomedical named entity recognition”, *Bioinformatics*, vol. 33, no. 14, pp. i37-i48.v, 2017.
- [18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, “Distributed representations of words and phrases and their compositionality”, *Adv. Neural Inf. Process. Syst.*, pp. 3111-3119, 2013.
- [19] S. Hochreiter, J. Schmidhuber, “Long short-term memory”, *Neural Comput.*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [20] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck, “The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions”, *J. Biomed. Inform.*, vol. 46, no. 5, pp. 914-920, 2013.
- [21] I. Segura-Bedmar, P. Martínez, M. Herrero Zazo, “Semeval-2013 Task 9: Extraction of drug-drug interactions from biomedical texts (DDIextraction 2013)”, in *Proc. SemEval*, vol. 2, pp. 341-350, 2013.
- [22] R. Chalapathy, E. Z. Borzeshi, M. Piccardi, “An investigation of recurrent neural architectures for drug name recognition”, *arXiv preprint arXiv:1609.07585*, 2016.