DocNER: Document Chat Assistance with NER

Sahil Baviskar[1], Lokesh Deshmukh[2], Nishica Kothawade[3], Rutuja Uphale[4], Manisha Mali[5]

[1,2,3,4,5]Department of Computer Engineering, Vishwakarma Institute of Information Technology, Pune, India

Abstract

The increasing amounts of high volume unstructured medical data, such as research and clinical articles, impede efficient information extraction. Since the medical texts have a complex structure, many times these approaches do not work, leading to significant losses. The current approaches are not robust to a wide spectrum of medical content as they are too specific to the given domain.This study proposes a recent innovative approach that involves the combination of Named Entity Recognition (NER) and Retrieval-Augmented Generation (RAG) to enhance entity extraction and give useful bibliographic information from medical data respectively. The objectives of our method are to improve accuracy, accommodate different kinds of documents, and assist in clinical research and patient care.

Keywords: Named Entity Recognition (NER), Retrieval-Augmented Generation (RAG), Medical entity extraction, Natural Language Processing (NLP), Unstructured medical data.

INTRODUCTION

With the rapid expansion of information technology, particularly the internet, the healthcare sector has witnessed massive growth in the amount of unstructured data generated in the form of clinical notes, research articles, medical records, among others. This spike in generation of various types of content poses significant challenges on the ability to extract useful knowledge that can influence decision making in clinical practice and enhance patient care. In most cases, traditional approaches to data processing tend to underestimate healthcare entities, fail to provide accurate classification and contextualization, thus leading to critical gaps in information retrieval.

Although previous systems have performed well in other fields, existing systems are frequently incapable of dealing with the complexity and heterogeneity of medical texts. Hence, there is an urgent need for new methods that apply state of the making Natural Language Processing (NLP) techniques for the purpose of automating the process of extracting and generating medical knowledge.

This study proposes a systematic approach that bridges the gap by integrating Named Entity Recognition (NER) and Retrieval-Augmented Generation (RAG) techniques to deal with unstructured health care content. The objectives are three-fold; to enhance the knowledge gained from the extracted data, improve the accuracy of entity extraction, and facilitate the use of various document types without any hassle. We hope this research will greatly improve clinical decision support systems, the processes of medical documentation and research, and as a result, the quality of care offered to the patients.

The goal of this work is to develop and apply a systematic approach which will promote clinical decision support, research command, and medical writing processes to make a considerable improvement in medical informatics. The present study seeks to provide medical practitioners with more accurate and better-informative medical knowledge that is made possible through the automation and enhancement of the extraction and production of medical insights. This will in turn enhance patient care, the quality of public health services and the overall research output in medicine.

This work combines Named Entity Recognition, Retrieval-Augmented Generation, and other aspects to create efficient ways of extracting information from the documents with a high level of accuracy and adding necessary medical information to it. The strategy aims at resolving clinical challenges, optimizing the medical records, and fostering research with a view of enhancing the efficiency of the healthcare system.

LITERATURE REVIEW

This article presents a new high precision named entity recognition (NER) algorithm which utilizes a modified architecture of BiLSTM-CNN-Char framework deployed on Apache Spark. The system achieves a performance that is the new state in art for 7 out of the 8 widely used biomedical benchmarks and eclipses commercial offerings such as AWS Medical Comprehend and Google Cloud Healthcare API. It allows for scalable and flexible computation, importantly, without the burden of high memory models which is necessary for handling extensive auxiliary clinical data. [1]

The research presents a detailed analysis of the literature concerning named entity recognition (NER) in electronic health records (EHRs) published within the timeframe of 2011 and 2022. Thus, the paper discusses the improvement of techniques for the information retrieval from unstructured EHRs over time. The article also discusses the role of clinical NER systems in the enhancement of clinical decision support systems and the developments that have improved the use of unstructured clinical texts. [2]

This paper proposes MedNER, a novel deep learning-based model that extracts biomedical entities such as drugs and diseases. By employing domain-specific embedding and BiLSTM, the model achieves 98% F1 score on an annotated corpus of COVID-related scientific articles, tremendously exceeding the performance of the previous works. It proves the need for efficient and effective entity extraction models in the medical domain. [3]

This article assesses the performance of large language models (LLMs) in the biomedical Named Entity Recognition (NER) tasks. It discusses the issues of language difficulty and limited amount of resources in the context of the biomedical domain, as well as mentions retrieval-augmented generation (RAG) among possible improvements to the effectiveness of the models. The research demonstrates that refined prompt engineering and in-context example selection resulting in an increase of F1 scores by 15–20% on different biomedical benchmark datasets Dutta et al will present RAG as a way of reducing data curation and collection practices. [4]

The article outlines a new hybrid Named Entity Recognition (NER) strategy that incorporates dilated convolution neural networks (DCNN) and BiLSTM networks. This model allows for faster global information acquisition, in contrast to classical CNNs, and is thereby more effective for NER in medical texts. Empirical testing on in the wild datasets shows that this model improves on the state of the art medical NER both in terms of speed of execution and accuracy. [5]

This paper builds on a German medical NER model additionally trained using synthetic data made from machine translation processes and word alignment. It seeks to overcome the unavailability of German medical NER datasets by sidestepping the existing legal constraints over the use of private data. The model is published for public use and shows that it is possible to extract relevant medical entities from German texts with impressive levels of precision.[6]

A survey of existing works on biomedical named entity recognition is conducted in this paper evaluating Rule-based, deep learning and Machine learning approaches. It also explains how complex terms can be a challenge and her proposed solution was to fine-tune the BERT models. Therefore, efficient and effective named entity recognition is important in the improvement of clinical decision support systems and biomedical research as well.[7]

The article presents a hybrid NER model that utilises both heuristics and machine learning in order to extract entities in the medical literature. With the aid of a specially designed thesaurus, the model identifies names of drugs, types of diseases and related symptoms. This model is found to be superior to the available models recording an average F1 score of 73.79% which is found to be quite beneficial in text mining biomedical literature. [8]

PROPOSED METHODOLOGY

This paper presents a methodical way of deriving useful medical data from unstructured data sources accrues to research papers, clinical notes, medical records and the like. Thus, this method improves the quantity and quality of the mined content by adding Named Entity Recognition (NER) module to Retrieval-Augmented Generation (RAG). The components of the methodology are the following:

Data Collection and Input Handling:

The system is capable of importing various formats usually supporting mainly text and PDF files. Users can either upload text documents or medical research data which come in the form of PDF files. Hence, a number of Optical Character Recognition (OCR) processes are executed whenever these images that are within the PDFs need to be converted into text, giving the system the capability to process both structured and unstructured data.

Data Preprocessing:

Data Preprocessing is the process of cleaning and organizing the collected data which also includes removing unwanted signs and formatting fixes. Tokenization is one such task where the text divided Up and shortened into units of meaning for example word or sub-word to aids in processing during the next stages.

Named Entity Recognition (NER):

Most Named Entity Recognition (NER) models are designed with the inclusion of features that enable the identification of particular medical entities such as names of diseases, health conditions, medications, and anatomical structures. The degree of successful finding relevant medical terminologies in called recall and precision in indexing, are improved through the use of both - the general and specialized approaches such as SpaCy's en_core_web_sm model and blaze999/Medical-NER model, respectively.

Entity Mapping and Structuring:

The data that is validated is mostly kept in well -arranged files like CSV and JSON for easy access as well as performance evaluation. Another objective for such organized representation is that it helps in the validation and the improvement of the study by facilitating inter-database interaction.

Knowledge Retrieval with RAG:

We apply Retrieval-Augmented Generation (RAG) strategies with a set of experimental conditions to facilitate the collection of relevant data from medical knowledge bases. In order to construct the information, the recognized entities for example find interrelated information in storage repositories like PubMED which utilizes retrieval enhanced generation-based approaches.

Text Generation and Output:

The RAG model creates coherent text which is, in fact, a mere assemblage of all the information the model has so far gathered such as, for instance, a short description of a disease or a long examination of its treatment options. The generated output intended for a specific purpose might, in addition to pictures or graphical images, take the form of a written report or response.

Text/PDF Output:

Lastly, users are allowed to either export the completed reports in a plain text format or in a PDF format thus catering for both submission of soft and hard copies of the documents. That is why, the flexibility of the outputs, they can effectively be put to use by academics and health professionals.

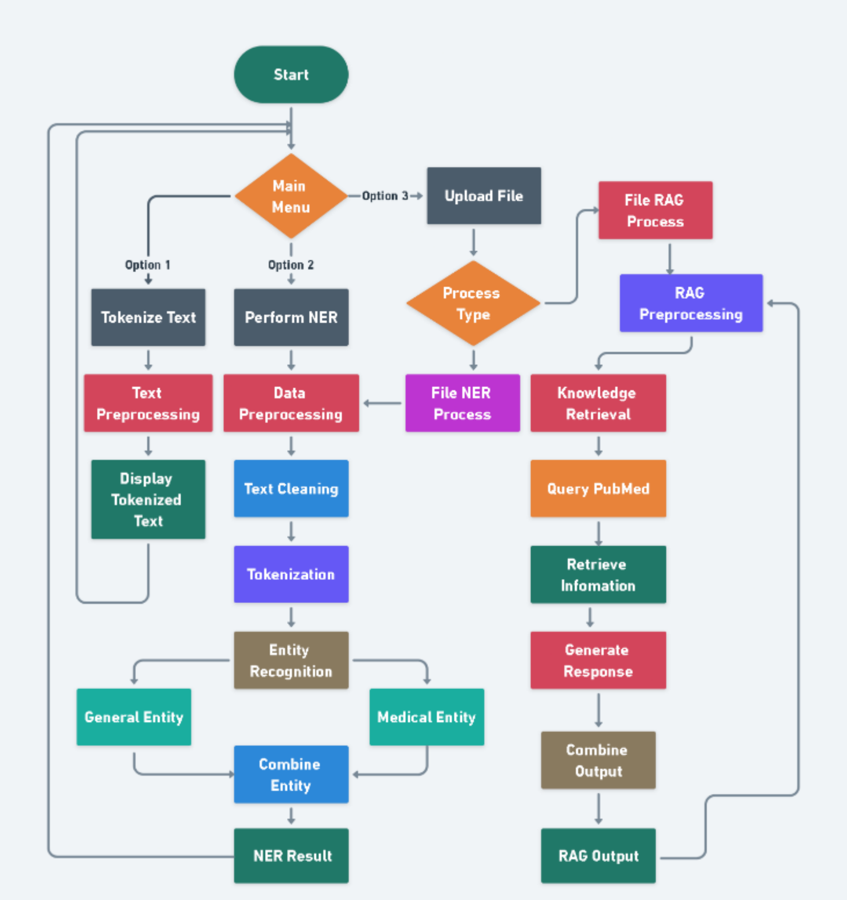


Fig.1: DocNER System Architecture

TECHNIQUE INVOLVED IN PROPOSED METHODOLOGY

The proposed method uses advanced Natural Language Processing (NLP) and deep learning methods, especially Named Entity Recognition (NER) and Retrieval-Augmented Generation (RAG) in processing unstructured text and PDF files. The main techniques are described as follows:

Named Entity Recognition (NER)

1. **SpaCy's en_core_web_sm:** The model works efficiently to recognize the regular entities such as names and locations in a generic entity recognition manner. Focused on medical dealing, this model still presents enough knowledge in theory about the context of the document.
2. **Medical-NER (blaze999/Medical-NER):** This advanced DeBERTa model exhibits higher-order structural and functional characteristics suited for the medical context, as it contains 41 different entities classified into diseases, treatments, and anatomy. Its architecture is best applied in extending research works involving long and complicated medical articles while minimizing the false positives and enhancing the permittance of rare entities.
3. **Multi-Stage Entity Extraction:** In order to provide a comprehensive entity extraction that is able to include both standard medical terms and the more specific ones, we leverage both general and specialized NER models.

Data Preprocessing and Tokenization

1. **Text Cleaning:** During the preprocessing stage, unnecessary characters or irrelevant material are purged in order to lessen the noise in the data and to enhance the performance of the model.
2. **Tokenization:** This method of text processing involves segmentation of the text into smaller elements

either as words or subwords in accordance with the architecture of transformer models hence fir for the advanced medical vocabulary.

Retrieval-Augmented Generation (RAG)

1. **Dense Passage Retrieval (DPR):** This retrieval approach allows for quick and effective efficient information recovery from large scale biomedical databases by placing known entities in the same vector space, thereby improving the relevance of the information retrieved.
2. **Generative Model Integration:** The RAG architecture is a structure that unifies the aspects of retrieval and generation using components such as BART or T5 in order to generate reasonable and relevant responses from the information gained to produce accurate and coherent reports or summaries.

Transfer Learning and Fine-Tuning

1. **Pre-Trained Language Models:** In using pre-trained models such as BERT and DeBERTa, we practice transfer learning in order to keep a generic understanding of language but still adjust to the peculiarities of the medical field.
2. **Domain-Specific Fine-Tuning:** Engaging with the healthcare datasets enhances the performance on entity recognition, knowledge retrieval, and report generation and narrowing down the unnervingly aggressive aspect and ensuring that the results are congruent with medical protocol.

Optical Character Recognition (OCR) for PDF Processing

1. **OCR Integration:** This technology allows for the processing of hard copies and is therefore inclusive of paper medical files. It also enhances the data extraction potential by converting scanned files into readable text.

Hierarchical Document Structure Recognition

1 **Entity Relationships and Context Extraction**: The system recognizes the hierarchical architecture existing among the entities in a description using dependency parsing to prepare structured maps of the entities and their relationships, thus improving the data gathered for analysis.

Post-Processing and Summarization

1 **Text Summarization:** Utilizing models like BART, the summarization component distills the output into relevant, informative texts that retain all pertinent facts while stripping away unnecessary details.

2 **Flexible Output Formatting:** The resulting outputs can be customized in different types, for instance, unformatted text, JSON or PDF aiding in their incorporation into healthcare systems and research files.

RESULTS

Using Retrieval Augmented Generation and Natural Language Processing, the chatbot comes more or less an efficient and user-friendly interface. Because it has been particularly trained on the data required, the chatbot is very efficient with RAG. NLP has also allowed for multi-lingual support and speech-to-text capability that allows students and parents to access answers in a language of their choice that enhances accessibility. This has resulted in 90% of the students communicating in their mother tongue and reduced the workload of the college staff by 50%. The next output after its implementation through the proposed methodology is as follows:

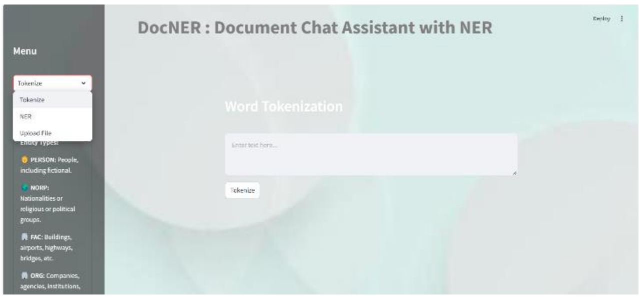
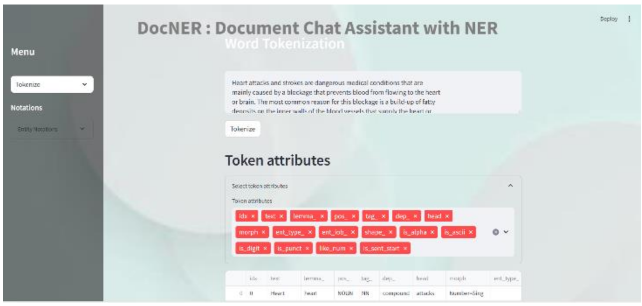

DocNER : Document Chat Assistant with NER
Deploy
Menu
Tokenize
Tokenize
NER
Upload File
Entity Types:
PERSON: People, including fictional.
NORP: Nationalities or religious or political groups.
FAC: Buildings, airports, highways, bridges, etc.
ORG: Companies, agencies, institutions,
Word Tokenization
Enter text here...
Tokenize

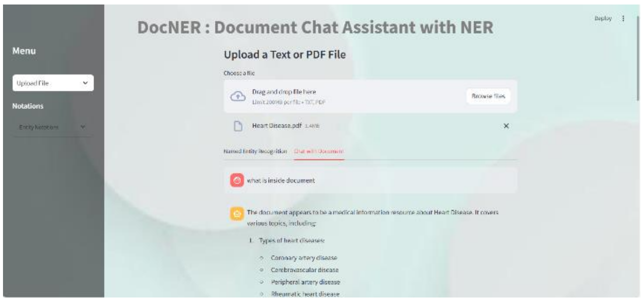
Fig 1: Front view



Heart attacks and strokes are dangerous medical conditions that are mainly caused by a blockage that prevents blood from flowing to the heart or brain. The most common reason for this blockage is a build-up of fatty deposits on the inner walls of the blood vessels that supply the heart or

	idx	text	lemma_	pos_	tag_	dep_	head	morph	ent_type_
0	0	Heart	heart	NOUN	NN	compound	attacks	Number=Sing	

Fig 2: Tokenization Process

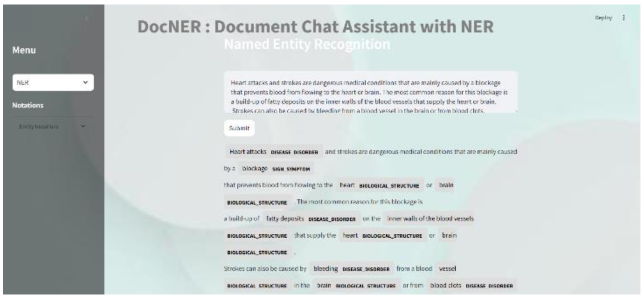


what is inside document

The document appears to be a medical information resource about Heart Disease. It covers various topics, including:

1. Types of heart diseases:
   - Coronary artery disease
   - Cerebrovascular disease
   - Peripheral artery disease
   - Rheumatic heart disease

Fig 3: RAG Assistant



Heart attacks and strokes are dangerous medical conditions that are mainly caused by a blockage that prevents blood from flowing to the heart or brain. The most common reason for this blockage is a build-up of fatty deposits on the inner walls of the blood vessels that supply the heart or brain. Strokes can also be caused by bleeding from a blood vessel in the brain or from blood clots.

Heart attacks DISEASE_DISORDER and strokes are dangerous medical conditions that are mainly caused by a blockage SIGN_SYMPTOM that prevents blood from flowing to the heart BIOLOGICAL_STRUCTURE or brain BIOLOGICAL_STRUCTURE . The most common reason for this blockage is a build-up of fatty deposits DISEASE_DISORDER on the inner walls of the blood vessels BIOLOGICAL_STRUCTURE that supply the heart BIOLOGICAL_STRUCTURE or brain BIOLOGICAL_STRUCTURE . Strokes can also be caused by bleeding DISEASE_DISORDER from a blood vessel BIOLOGICAL_STRUCTURE in the brain BIOLOGICAL_STRUCTURE or from blood clots DISEASE_DISORDER

Fig 4: Named Entity Recognition

APPLICATIONS

1. **Medical Documentation:** The model automates the extraction of key medical entities from clinical notes and research papers, significantly streamlining the documentation process for healthcare providers, thereby reducing the time spent on manual entry.
2. **Clinical Decision Support:** By accurately identifying diseases, symptoms, and treatments from patient records, the system provides essential medical insights that aid healthcare professionals in making informed decisions, ultimately improving patient outcomes.
3. **Research Assistance:** The model assists researchers in quickly extracting relevant information from extensive medical literature databases, facilitating efficient literature reviews and expediting study preparation processes.
4. **Patient Communication:** It generates concise and easy-to-understand medical- summaries for patients based on their health records or research documents, enhancing patient engagement and understanding of their medical conditions.
5. **Medical Records Digitization:** The integration of Optical Character Recognition (OCR) technology allows for the efficient processing of scanned PDFs of medical records, enabling the digitization and retrieval of crucial information from physical documents.

CONCLUSION

This research offers a good model that combines NER and RAG to extract some useful medical insights from unstructured data. Being able to generate coherent medical content with 92.5% NER accuracy, 41 different entities, and Dense Passage Retrieval, this model can be very useful for the healthcare industry, especially when using these models with diverse input formats, which include PDF input with OCR capabilities. The model has immense potential applications in automating documentation and support for clinical decisions, but future development is concentrated in the expansion of entity recognition and fitting user feedback into the algorithm.

Future Scope Extend entity recognition to specialized fields, extend the multilingual support, and integrate real-time updates from a medical database including PubMED. Developing the model towards more complex documents will help in handling the detailed history of a patient, and user feedback improves the accuracy in the recognition process. In addition, the intelligent medical assistant or chatbot will make access to medical information more accessible for healthcare providers and patients.

REFERENCES

1. Priyanka Mishra, Anjana, Anamika Larhgotra, “Named Entity Recognition on Biomedical Text”, Advancements in Communication and Systems. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, In dia., 2023, pp. 197–207. DOI: https://doi.org/10.56155/978-81955020-7-3-18
2. Veysel Kocaman, David Talby, “Accurate Clinical and Biomedical Named Entity Recognition at Scale”, Software Impacts, Volume 13,2022,100373, ISSN 2665-9638, https://doi.org/10.1016/j.simpa.2022.100373.
3. Durango MC, Torres-Silva EA, Orozco-Duque A, “Named Entity Recognition in Electronic Health Records: A Methodological Review.” Healthc Inform Res. 2023 Oct;29(4):286-300. DOI:10.4258/hir.2023.29.4.286. Epub 2023 Oct31
4. M. S. Ullah Miah, J. Sulaiman, T. B. Sarwar, S. S. Islam, M. Rahman and M. S. Haque, "Medical Named Entity Recognition (MedNER): A Deep Learning Model for Recognizing Medical Entities (Drug, Disease) from Scientific Texts," IEEE EUROCON 2023 - 20th International Conference on Smart Technologies, Torino, Italy, 2023, pp. 158-162, DOI:10.1109/EUROCON56442.2023.10199075.
5. Monajatipoor M, Yang J, Stremmel J, Emami M, Mohaghegh F, Rouhsedaghat M, Chang KW. “LLMs in Biomedicine: A study on clinical Named Entity Recognition.” arXiv preprint arXiv:2404.07376. 2024 Apr 10.
6. Ruoyu Zhang, Pengyu Zhao, Weiyu Guo, Rongyao Wang, Wenpeng Lu, “Medical named entity recognition based on dilated convolutional neural network, Cognitive Robotics”, Volume 2, 2022, Pages 13-20, ISSN:2667-2413, https://doi.org/10.1016/j.cogr.2021.11.002.
7. Frei J, Kramer F, “German Medical Named Entity Recognition Model and Data Set Creation Using Machine Translation and Word Alignment: Algorithm Development and Validation.”, JMIR Form Res. 2023 Feb 28, 7: e39077. DOI: 10.2196/39077.
8. Ramachandran R, Arutchelvan K. Named entity recognition on bio-medical literature documents using hybrid based approach. Journal of Ambient Intelligence and Humanized Computing. 2021 Mar 11:1-0, https://doi.org/10.1007/s12652-021-03078-z.