International Journal of Computational Intelligence in Control

Natural Language Processing For Medical Information Extraction

B. Backiyalakshmi and M. Rajakumar
Jamal Mohamed College, Trichy, (Affiliated to Bharathidasan University, Tiruchirappalli), India

Received: 28th September 2021
Revised: 20th October 2021
Accepted: 19th November 2021

Abstract: Extracting medical data from narrative clinical papers using NLP is called NLP-based retrieval of medical information. In many cases, retrieving data in this manner may be really beneficial. This study examines and discusses the use of Natural Language Processing (NLP) to extract medical issues from clinical records. In addition to the architecture of the system that has been proposed, this article discusses the methods employed in this sector. Medical data extraction, on the other hand, is challenging due to the difficulty in recognising symptoms and illnesses. The suggested expert systems would analyse the input, which may be a question about disease or a list of symptoms, and then make an attempt to offer the proper response. Data processing, query processing, data extraction, response matching, and a user interface are all included in the proposed system. Traditional rule-based systems do not have access to this information, therefore instead we may utilise natural language processing (NLP) to obtain this information, which may give us with the answers we are looking for inthe medical field.

Keywords: Medical data extraction, narrative text, and natural language processing

1. INTRODUCTION

Free text in electronic health records must be interpreted, and this is a difficult step to take given the increasing usage of electronic health records and the subsequent interest in quality improvement and research created by these records. Another major source of data is the biomedical literature, which would substantially benefit from the implementation of an organisation based on narrative language. There are a slew of approaches for gleaning knowledge from medical books. Noun entity recognizers, resolvers for cross-references and part of speech tags are only some of the tools used in the natural language processing method. As a result of the difficulty of medical terminology, standard textbooks require simpler tools, whereas medical textbooks demand more complex ones.

Motivation

The statements that follow express the necessity of a computerised programme.

- **Processing of Text is Necessary:** As long as we continue to work with data, we need to always be able to access it by means of a format known as a database. We are squandering a lot of information merely due to the fact that it is not presented in the appropriate style. It is required that this data be processed by a system, as indicated in the statement. That way, we won't ever have to worry about not having enough data, and we'll be able to make better use of the information that's currently available to us in the form of free text.

- **Medical Text Processing is Required:** Within the realm of medicine, one may easily have access to a wealth of information. However, the application of Natural Language Processing in the field of medicine has not advanced nearly as far as it has in other fields. Electronic Medical Records, often known as EMRs, as well as several other types of medical data are easily accessible. Data can be extracted from these medical records, but only by a select few computer systems. As a consequence of this, the aforementioned data has to be processed, and natural language processing can be of assistance in this endeavour.
- **Diagnosis of Diseases Requires an Automated System:** The typical patient has no means of knowing what kind of diagnosis or treatment he will receive until after he has seen the attending physician. It is concerning that some individuals are incapable of recognising the early warning signs of any disease. Patients who have access to an automated system for the diagnosis of their ailment will, at the very least, be able to judge how severe their condition is.
- **Doctors require a Computerised System:** The ability to remember a significant amount of information is essential for a career in medicine. If they could utilise some kind of automated system to help them diagnose a patient's disease, that would be a terrific idea. It also notes that medical professionals, such as physicians, require an automated system.

It is necessary to develop a system that is able to analyse medical documents in order to produce a diagnosis of the disease and the degree to which it has progressed. Currently, we are putting to the test a method to risk categorization for general disorders that is based on Natural Language Processing (NLP). This technique uses medical data. In the long run, researchers expect that by providing patients with information about their diseases and the severity of those illnesses, they will be able to enhance patient safety.

Retrieval of Data is Made Easier by Utilising NLP:

Rule-based systems are utilised by the overwhelming majority of computer programmes in order to retrieve information. In rule-based systems, there is a limit on the total number of rules that may be implemented. The quantity of information that may be obtained is constrained due to the fact that there are only a few rules to follow. Because it does not have any rules, the rule-based system is unable to give the required new data when itis called for. Natural language processing has the ability to make use of free-form text. The volume of freely accessible text data has ballooned to tremendous proportions in recent years. On the other hand, nobody ever makes use of it. The information that is contained in documents such as Electronic Health Records may be utilised in a variety of different ways. The unrestricted text provides details on several diseases, including their manifestations and the factors that contribute to their development. The use of natural language processing might be utilised in order to extract all of this information (NLP). One such finding is that medical professionalsdo not always use their entire skills while treating patients with various ailments. Clinicians may be able to improve the accuracy of their disease diagnosis by utilising a data extraction approach that is automated.

2. SIMILAR WORK

MedlEE, MetaMap, and the Linguistic String Project [9] are examples of medical extraction systems. MedLEE can be used to extract, organise, and encode clinical data, resulting in textual patient reports. The MedLEE system was developed by Columbia University's Biomedical Informatics Department, Columbia University's Radiology Department, and Queens College's Computer Science Department, under the direction of Carol Friedman. The National Library of Medicine's Dr. Alan Aronson developed MetaMap, a highly customizable piece of software (NLM). Metathesaurus ideas that are quoted in text can be converted to the UMLS Metathesaurus using this tool [9]. Between 1960 and 2005, the Linguistic String Project (LSP) was being developed in the field of computer language processing [9]. As a result, Zellig Harris's string theory, transformation analysis and sublanguage grammar were employed in the book's writings.

A wide range of techniques are employed in natural language processing, including relationship extractors, pos- taggers, co-ref resolutions, and NER, to name a few. Named Entity Recognition and Classification was invented by Branimir Todorovic and Svetozar R. Rancic by applying a context-based Hidden Markov Model. [14] Mauricio Hashem suggested the construction of a supervised approach for extracting named items from medical literature. Hidden Markov Model-Based System by Louise Deleger effectively adjusted Andreea Bodnari's Biomedical Named Entity Recognizer [14].

It is referred to in [15] that Jiaping Zheng devised a strategy for resolving clinical narrative coreferences. Clinical Relationships techniques were established by using patient narratives by Wafaa Tawfik and Abdel- moneim [7]. Annotated co-reference pairings total 7214 in the Ontology Development and Information Extraction corpuses, which have been used to create the relationships. Semantic, syntactic, and surface aspects can be trimmed from classifier training by using feature selection. In terms of working with noun phrases, relative pronouns and personal pronouns. It's feasible to construct machine learning using support vector machines, decision trees, and perceptrons with linear and radial basis function kernels [10].

It is referred to in [15] that Jiaping Zheng devised a strategy for resolving clinical narrative coreferences. Clinical Relationships techniques were established by using patient narratives by Wafaa Tawfik and Abdel- moneim [7]. Annotated co-reference pairings total 7214 in the Ontology Development and Information Extraction corpuses, which have been used to create the relationships. Semantic, syntactic, and surface aspects can be trimmed from classifier training by using feature selection. In terms of working with noun phrases, relative pronouns and personal pronouns. It's feasible to construct machine learning using support vector machines, decision trees, and perceptrons with linear and radial basis function kernels [10].

As a result of this method, semantic links may be drawn between words and phrases.

(i) Medical entities are recognised as such.

(ii) Each pair of entities must be linked in the proper semantic way.

The first step towards attaining this aim is to increase the use of metamaps. The following stage is based on semi-automatically created patterns of language from the previously selected corpus. Using semantic criteria, anevaluation of the treatment-disease link may be made.

3. INFORMATION RETRIEVAL

In the 1960s, a subfield of Artificial Intelligence and Linguistics known as "Natural Language Processing" (NLP) was formed in order to investigate issues in the automatic synthesis and interpretation of natural language. This was done with the goal of improving machine translation. The following is a list of the three primary components that make up a text retrieval system:

1. Documents in the form of records
2. Indexer
3. Information retrieval methods.

Records can be recovered using extraction methods with the aid of an indexer. Any or all of these processes can be supplemented by Natural Language Processing. Both the query and the document are interpreted and storedin NLP. The processes listed below should be followed if NLP information retrieval is included.

Step 1: Preparation of Records.**Step**
2: processing of Queries **Step 3:**
Query retrieval.
Step 4: Sorting and Ranking

4. PROPOSED WORK

In addition to document and query processing using natural language processing, these are the two core components of the proposed system. The successful completion of phases one and two of the project is essential. The architecture illustrates the five components that come together to form the suggested system. In addition to that, it features a knowledge database. The knowledge base contains copies of all of the patients' medical histories:

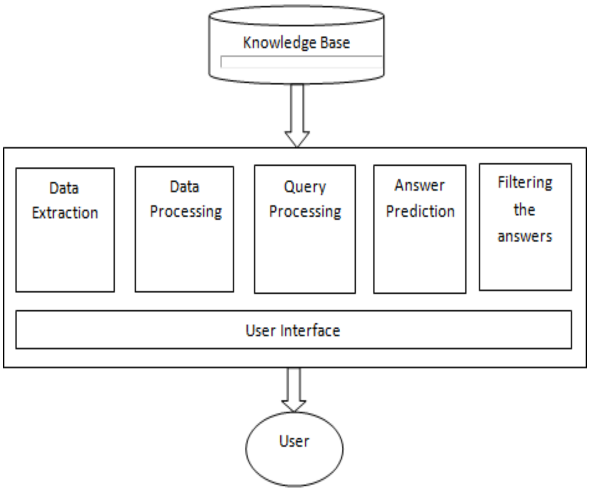


Fig. 1: The Proposed System's Architecture

Flow of the Proposed System:

Request will be sent to the module responsible for query processing, which will then process it. The query willbe mined for valuable information. They say papers will be analysed in order to get the right replies from them.

Question and answer matching aims to discover the most appropriate response from a pool of potentials The following modules are included in the proposed system:

1) Extraction of Data
2) Processing of Data
3) Retrieval of Queries
4) Congruency in Reply
5) Answers are Sorted

Extraction of Data

The Knowledge Base will be managed by this module. A search for information on the Internet will be carried out. Biomedical books on a variety of systems will be included in the knowledge base. It can be filled with any sort of free text that includes information on a certain ailment.

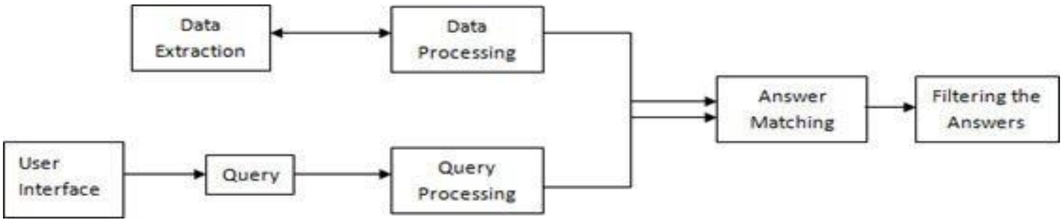


Fig. 2: The natural progression of the project.

Processing of Data

During the processing of documents, a Natural Language Processing (NLP) system is utilised. The method will consist of a number of steps, some of which include tokenization, relationship extraction, section splitting, and others.

Retrieval of Queries

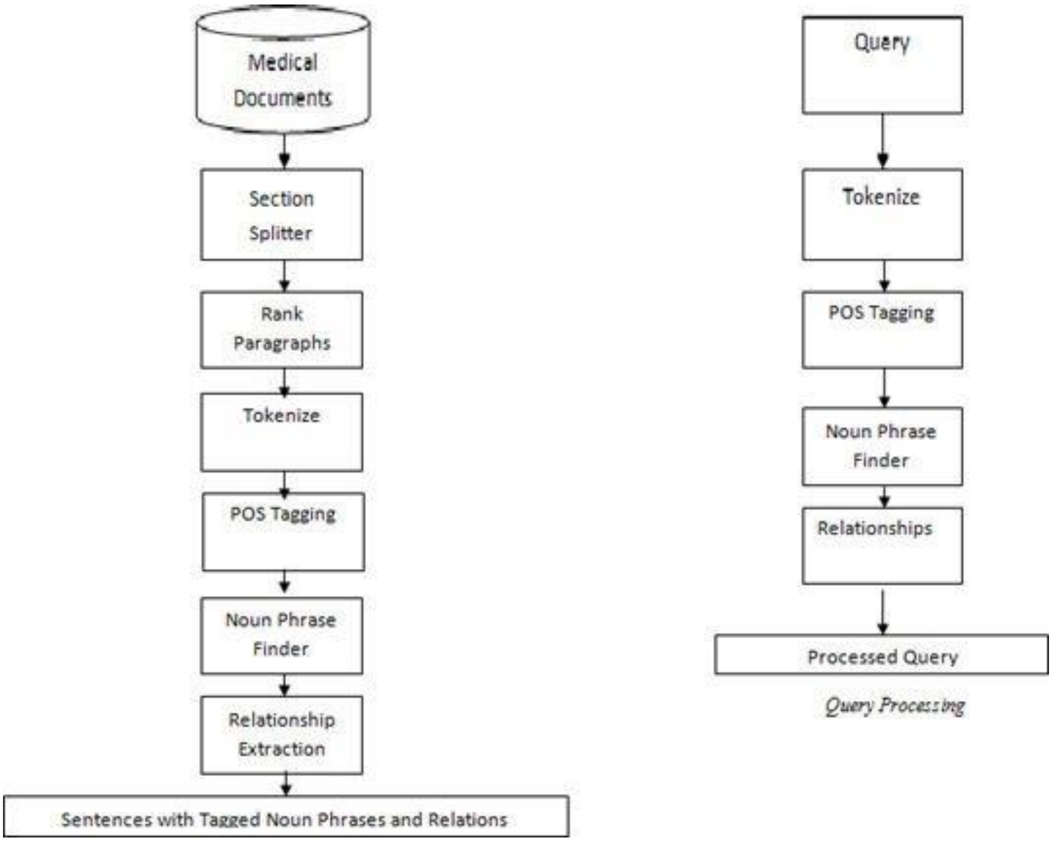
Natural language processing (NLP) will be used to aid with query processing. Extraction of significant linkages and keywords will be done using this tool

Congruency in Reply

Predicting an answer will be based on the relationships and keywords that are provided.

5. RESULTS

The project's goal is to provide the most accurate and reliable responses to disease-related questions. The system should be able to provide a plausible illness name based on a user-supplied list of symptoms.



Query Processing

Fig.3: Document Processing

6. CONCLUSION

The use of natural language processing (NLP) to the extraction of medical text is a topic that is now receiving a lot of attention. Because of the significant differences between medical language and other forms of writing, more advanced NLP methods are required. The extraction of text is another another use for many of these systems. Methods for extracting information from medical texts need to be refined. The current inquiry indicates that the proposed system is able to retrieve information regarding illnesses.

REFERENCES

1. Andreea Bodnari, Louise Deleger, Thomas Lavergne, “A Supervised Named-Entity Extraction Systemfor Medical Text”
2. James Freeman-Hargis “Introduction to Rule-Based Systems” at <http://ai-depot. com/ Tutorial/ RuleBased.html>
3. Ngô Thanh Nhàn “linguistic string project - medical language processor” at http:// www. cs.nyu. edu/ cs/projects/lsp/
4. The Brandeis University, “MedLEE” at <http://www.medlingmap.org/taxonomy/term/80>
5. Asma Ben Abacha, Pierre Zweigenbaum, “Automatic extraction of semantic relations between medical entities: a rule based approach” From Fourth International Symposium on Semantic Mining in Biomedicine (SMBM)
6. Dan Shen Jie Zhang Guodong Zhou, “Effective Adaptation of a Hidden Markov Model-based Named Entity
7. D. Nagarani, Avadhanula Karthik, G. Ravi, “A Machine Learning Approach for Classifying Medical Sentences into Different Classes”, IOSR Journal of Computer Engineering (IOSRJCE) Volume 7, Issue 5 (Nov-Dec. 2012), PP 19-24
8. Faguo ZHOU Enshen WU, “The Design of Computer Aided Medical Diagnosis System Based on Maximum Entropy” 978-1-61284-729-0111 2011 IEEE
9. Hinxton, UK. 25-26 October 2010
10. Jiaping Zheng,1 Wendy W Chapman,2 Timothy A Miller,1 Chen Lin, “A system for coreference resolution for the clinical narrative”, J Am Med Inform Assoc (2012). doi:10.1136/amiajnl-2011-000599
11. Kyle D. Richardson1, Daniel G. Bobrow1, Cleo Condoravdi1, Richard Waldinger2, Amar Das3, “English Access to Structured Data”, 2011 Fifth IEEE International Conference on Semantic Computing
12. Khan Razik, Dhande Mayur , “To Identify Disease Treatment Relationship in Short Text Using Machine Learning & Natural Language Processing”, Journal of Engineering, Computers & Applied Sciences (JEC&AS), Volume 2, No.4, April 2013
13. L. Smith1, T. Rindflesch2 and W. J. Wilbur, “MedPost: a part-of-speech tagger for bioMedical text”, Vol. 20 no. 14 2004, pages 2320–2321, bioinformatics/bth227
14. Lucila Ohno-Machado, Editor-in-chief, Prakash Nadkarni, Kevin Johnson “Natural language processing: algorithms and tools to extract computable information from EHRs and from the biomedical literature”, amiajnl-2013-002214
15. Romer Rosales, Faisal Farooq, Balaji Krishnapuram, Shipeng Yu, Glenn Fung, “Automated Identificationof Medical Concepts and Assertions in Medical Text Knowledge Solutions” , AMIA i2b2/VA text mining challenge