Relation Extraction in Clinical Text using NLP Based Regular Expressions

Veena G, Hemanth R, Jithin Hareesh
Department of Computer Science and Applications, Amrita Vishwa Vidyapeetham, Amritapuri, India
veenag@am.amrita.edu, hemanth16295@gmail.com , jithinhareesh@gmail.com

***Abstract*—This paper is about relation extraction system for medical field related data. The major aim of our work is to retrieve different medical related data and to find out the relation between extracted medical data. Medical data usually contains a lot of unstructured or semi-structured data, by implementing methods like labeling and path similarity analysis we are able to convert it into a structured or classified form. Other methods that we use in our work are web scrapping, regular expressions, and part-of-speech tagging. All these methods are implemented in python.**

Keywords: Relation extraction, Regular expressions, Part-of-Speech tagging, Path similarity, NLP

I. INTRODUCTION

Relation extraction(RE) task is used to obtain the relations between entities in clinical texts. It is similar to information extraction(IE) but in information extraction task only data or information is extracted and not the relationships between entities in data. These types of data extraction are done with the help of Natural Language Processing(NLP). This helps a computer program to understand or process human language there by making the processing task easier. There are many useful applications for natural language processing, it is used in different areas like artificial intelligence(AI), medical field etc. In natural language processing there are different approaches for machine learning , but we use an unsupervised machine learning approach in our work, these types of approaches does not use any labeled trained data. We know that medical related data usually contains a lot of information's which are unstructured or semi-structured which means these information's either does not have a predefined data model or not organized in a predefined manner, these information's are usually text heavy. This results in irregularities and ambiguities in the result, thus we may not get a proper result always and also the information's will be in the form of a machine readable document format and not in a human readable form which makes it difficult for anyone who wants to retrieve information's from it, So to obtain information's we have to use various Relation extraction(RE) tasks. These tasks or methods helps us to Process, extract and encode organized data from unorganized or semi-organized machine readable documents. Different methods we use in our project are Web-scrapping, Regular expressions, Part of speech(POS) tagging, Labeling related medical words and Path similarity analysis. Web scrapping is done to extract or load the contents of a website into our local database. Regular expressions is used to extract different medical related words,sentences or paragraphs from extracted data. Part-of-speech tagging is done to identify the singular and plural nouns from our data. In labeling part we give an appropriate related label to each word. Path similarity analysis is done to calculate the similarity and relations between a set of concepts , these pair of concepts are called syn-sets , all of which are based on the lexical database Word-Net. Word-net contains the syn-sets which helps in providing definition's and usage examples.

Section II explains the related works of our paper and the proposed system and methodology of this work are explained in section III. Results and conclusions are presented in section IV.

II. RELATED WORK

There have been several previous works on medical data extraction from clinical texts, some of those works are [1] [2] [3] [4] [5]

A. Mykowiecka, M. Marciniak and A. Kupsc [1] developed a data retrieval system which functions upon certain rules. This system extract information's from patients clinical data. This work was made for clinical text which are polish in nature. There are different types of applications developed in-order to choose information's from clinical documents. One is mammogram reports and the other is clinical records from people who is suffering from diabetes. Firstly, they developed a unique ontology which eventually had its idea translate into different models depicted as typed featured structures hierarchy, then they use exclusive information extraction grammars to examine medical documents.

X. Fu and S. Ananiadou [2] developed a system which helps to improve the retrieval of concepts from medical records. Important information's related to medical issues, tests reports and treatments results are usually included in the patients medical documents with the help of human language which makes it a big task for the computerized programmed systems, so in order to solve this issue they use concept extraction. In their work they put forward an entity identification system which uses machine learning,

978-1-7281-0283-2/19/$31.00 ©2019 IEEE

particularly information's are extracted from patients discharge summaries or progressive notes. They did not use any outside information source, they have identified different pre-operating and post-operating procedures. These procedures are true-casing. abbreviation-disambiguation and distributional thesaurus look-up.

W. T. Abdel-moneim, M. H. Abdel-Aziz, and M. M. Hassan [3] developed a system which uses clinical relationship extraction techniques to extract data from patient narratives. This is done by building a system for medical research, proof based health care and for genotype or phenotype information processing. Their work has different parts, one part is about selecting the relationships between clinically important entity which is present in the clinical texts and the other part deals with statistical machine learning approach which are used to retrieve relationships in clinical texts.

D. Demner-Fushman and J. Lin [4] developed a system which deals with a type of knowledge extraction which can be used for question answering. This system provides a special chance in exploring question answering which are complex in the clinical domain. In their work they try to ope-rationalize important characteristics of proof based medicine. From their analysis it is also clear that domain related knowledge can be used to retrieve pico elements as frames with the help of medical literature analysis and retrieval system.

B. Rink, S. Harabagiu, and K. Roberts [5] develops a system which retrieves relationships between different concepts in clinical texts. The major objective of their work is to identify the relationships between medical issues, treatment reports and test results in the electronic clinical records with the help of a supervised machine learning method. They recognizes relationships between concepts in clinical texts and also assign their appropriate semantic types with the help of single support vector machine classifier. The resources used in this work are word-net, Wikipedia and general inquirer.

III. METHODOLOGY

Fig.1 and Fig.2 explains the architecture of our proposed system. The major processes in our system are web scrapping , regular expressions, part-of-speech tagging , labeling and path similarity analysis.

A. Web scrapping

This method is also called as web harvesting or web data extraction. We can use this data scrapping method to extract data from any websites we want. This method can be used in two ways. One is, it can be used to enter the internet directly with the help of http protocol and the other by using a web browser. We have used this method to access the world wide web directly without using any web browsers. Here we use

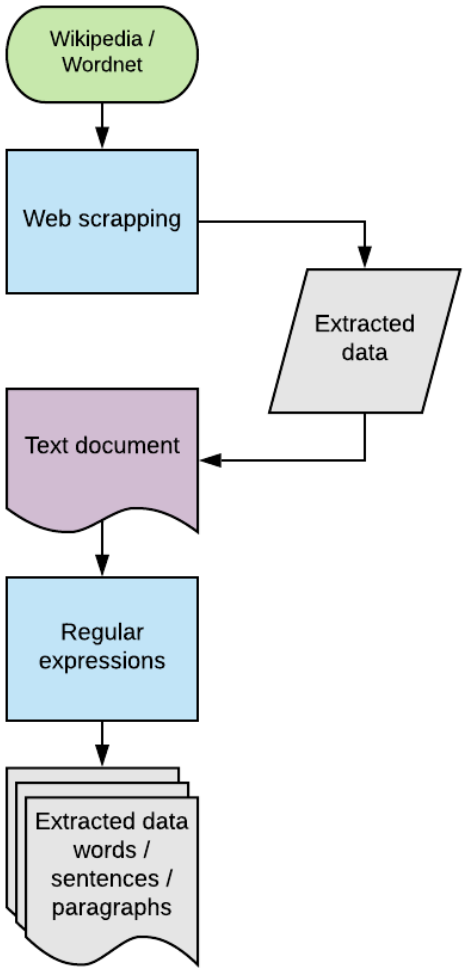

Fig. 1. Architecture Diagram

the web scraping technique manually as an automated process which is done or implemented using web crawler. That is it is simply a form of gathering the data that we require from the internet and it is typically stored or saved into a local database in our computer. Here we have saved the extracted data into a text document and it will be later used for data analysis or retrieval. The input data can be obtained from any sites in the internet. We have extracted a Wikipedia page about the disease cancer for our work.

B. Regular expression

These are unique series or set of characters which can be used to find or match other strings. In regular expression there is a specialized syntax which is enclosed in a pattern. In this syntax we can change the patterns in it which results in different types of data extraction according to the pattern we specify. In this work we use different types of regular expressions. Specifically we extract data in the form of words, sentences or paragraphs.

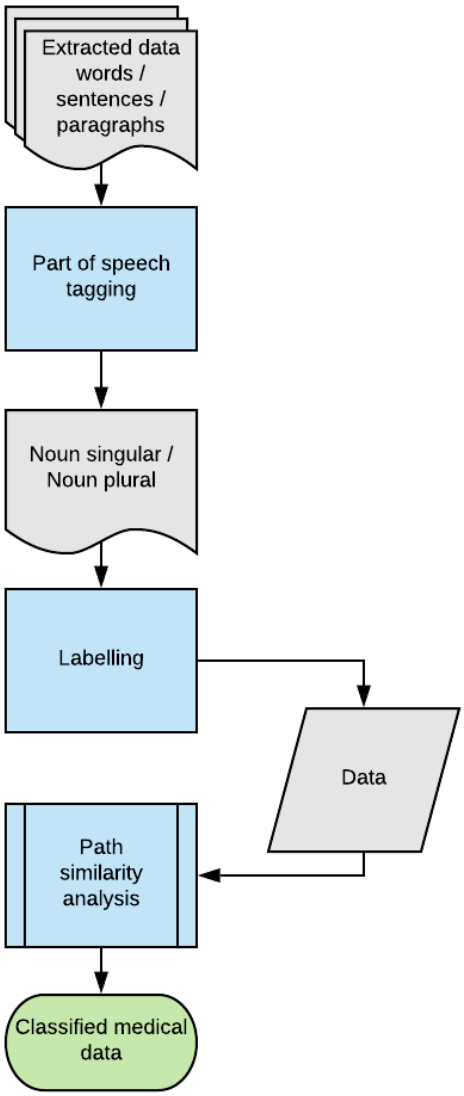

Fig. 2. Architecture Diagram

The basic regular expressions we used in our work are:

- re.find-all: This regular expression is used to find out all the matches in a single line of data.
- re.search: This regular expression is used to search for a particular pattern in a text or string.
- re.compile: This regular expression is used for compiling patterns into pattern objects.

In addition to that other regular expressions we created for our system are:

- re.search: This regular expression will return the first word after every "symptoms of" in our data

$$re.search(r'(.*)symptomsof(.*).*'$$

- re.compile: This regular expression we use will return every line in which the term "symptoms" are present

$$re.compile(''symptoms'', re.IGNORECASE)$$

- re.compile: This regular expression will return full paragraph after the word symptoms.

$$re.compile(r''symptoms*([\cdot]+|+)'')$$

C. Part of speech tagging

This is a method by which we mark up a word or a term in texts, sentences or corpus as matching to an appropriate part of speech. This is either regarding its definitions or its contexts with which it is related or adjacent words. After implementing POS tagging we get all the nouns, pronouns, verbs, adverbs, adjectives, conjunctions, prepositions, interjections in our data as the output. But instead of taking all the part-of-speeches, we take only the NN and NNS which stands for noun singular and noun plural. We take only the nouns because nouns contains more medically related or relevant words.

D. Labeling

In this method we use a list of medical related words which we obtain from the output of part-of -speech tagging and also add some additional medical words to the list with the help "Meta-Map" which is national library which contains data about medical related concepts. Then we give an appropriate label to each words in the list, So all the related words will be grouped together, that is for example the word lungs will be labeled or grouped with the word organ, the word bleeding will be labeled to the word symptoms , likewise the process will give a label to the entire list of words.

E. Path similarity

Path similarity is a type of analysis method which is done with the help of word-net hierarchy. Word-net is a big database of English. This database is lexical in nature, it contains nouns, verbs, adjectives etc, all of these are grouped into sets of synonyms which are called syn-sets, which we use to find the path similarity. For this method we take the list of words and its appropriate labels as the input. This analysis technique is defined by a metric which helps us to find similarities between different words. For example a word lung will compared to the word organ and the word disease, path similarity will make an analysis that which word is more related to the word lung. This is calculated using the metric values. Both the words will have different metric values towards the particular compared word, which comparison has a metric value more greater that word will be matched to that particular compared word.

INPUT: Medical related information from sites like Wikipedia or word-net

OUTPUT: Structured or classified medical data

Step 1: Extract words, sentences or paragraphs using different regular expression.;
Step 2: Apply part-of-speech tagging to the extracted data.;
Step 3: Identifies singular and plural nouns from POS tagging output .;
Step 4: Create a list of medical words which are obtained using POS tagging and also with the help of Meta Map medical library.;
Step 5: List of medical words are provided with appropriate labels .;
Step 6: Apply path similarity analysis, taking the list of medical words and their appropriate labels as input.;

Algorithm 1: Relation extraction in clinical texts.

IV. RESULT AND DISCUSSION

In our work we read web pages related to different diseases and extract different medical related information's from it with the help of methods like web scrapping and regular expressions. Web pages are read into our system using web scrapping method. Medical data are extracted from web pages in the form of words, sentences and paragraphs using different regular expressions.Different medical terms are identified with the help of part-of-speech tagging method in the form of noun singular and noun plural. As our output we get an appropriate label for each medical words we identified, for example a medical word lungs will be labeled to another medical word organ. We also implement path similarity analysis method as a sub method inside the labeling method. This particular analysis method helps us to find more similarities between different medical words which results in better classification of the medical data. For example this method can be used to compare a word lungs to other words like disease or organ, as a result we get which word is more closer to disease or organ by checking its path and that words will be labeled to the other word having the greatest path similarity value, here in this example the word lungs will be labelled to the word organ. So by implementing methods like labeling and path similarity analysis we are able to convert the unstructured or semi-structured medical data into structured or classified form. Likewise we are also able to identify different medical terms and also the relation between different medical terms. In addition to that we are also able to identify diseases for specific symptoms in our medical data.

We have scrapped a number of web pages related to various diseases and have identified the disease,symptoms,treatment and affected organ associated with that particular disease as our output.Our work gives a a well defined versions for doctors and other medically related people to get a quick understanding about a disease rather than reading a whole document.Our work is really a time saving one and it helps people to get a brief understanding about the diseases in a single view. Fig.3 shows a pictorial representation of our output.

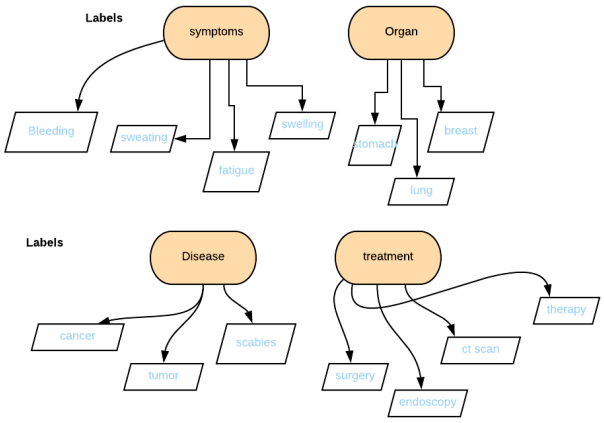

Fig. 3. Labelling of Terms

Fig.4 shows the variation of nouns corresponding to the number of sentences present in our extracted medical data. We get data in the form of sentences and nouns after implementing the regular expressions and POS tagging. In this figure the nouns also increases according to the increase in the number of sentences which shows the variation of medical data is in a linear way. By analyzing our work we find out that there are about 5276 sentences in our input data, which contains about 17900 noun, 2490 noun's, 8078 noun plural and 124 noun plural's.

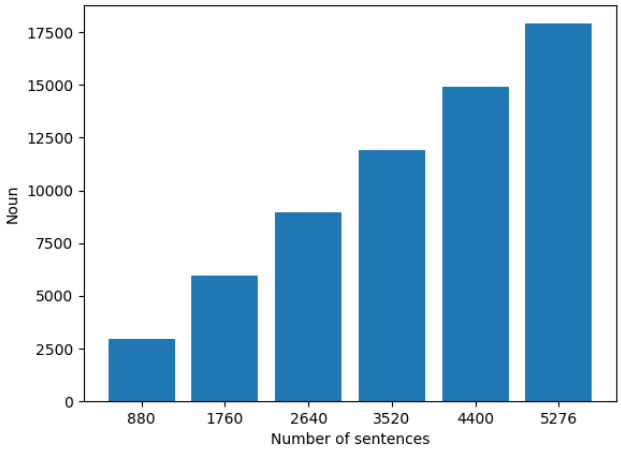

Fig. 4. Medical data variation

V. CONCLUSIONS

Our work can be used by medical personnel and also by common people. Our system can also be extended by adding

more functionality, for example we can extend our system for finding composition of medicines and their alternatives with the help of more advanced algorithms because a more advanced functioning system can deliver highly or more accurate data, so there is a greater scope for our work in the future.

REFERENCES

[1] A. Mykowiecka, M. Marciniak, and A. Kupść, "Rule-based information extraction from patientsâĂŹ clinical data," *Journal of biomedical informatics*, vol. 42, no. 5, pp. 923–936, 2009.

[2] X. Fu and S. Ananiadou, "Improving the extraction of clinical concepts from clinical records," *Proceedings of BioTxtM14*, pp. 47–53, 2014.

[3] W. T. Abdel-moneim, M. H. Abdel-Aziz, and M. M. Hassan, "Clinical relationships extraction techniques from patient narratives," *arXiv preprint arXiv:1306.5170*, 2013.

[4] D. Demner-Fushman and J. Lin, "Knowledge extraction for clinical question answering: Preliminary results," in *Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains*. AAAI Press (American Association for Artificial Intelligence) Pittsburgh, PA, 2005, pp. 9–13.

[5] B. Rink, S. Harabagiu, and K. Roberts, "Automatic extraction of relations between medical concepts in clinical texts," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 594–600, 2011.

[6] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn *et al.*, "Clinical information extraction applications: a literature review," *Journal of biomedical informatics*, vol. 77, pp. 34–49, 2018.

[7] S. N. Kim, D. Martinez, L. Cavedon, and L. Yencken, "Automatic classification of sentences to support evidence based medicine," in *BMC bioinformatics*, vol. 12, no. 2. BioMed Central, 2011, p. S5.

[8] J. D. Patrick, D. H. Nguyen, Y. Wang, and M. Li, "A knowledge discovery and reuse pipeline for information extraction in clinical notes," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 574–579, 2011.

[9] B. Tang, Y. Wu, M. Jiang, Y. Chen, J. C. Denny, and H. Xu, "A hybrid system for temporal information extraction from clinical text," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 828–835, 2013.

[10] M. G. Seneviratne, T. Seto, D. W. Blayney, J. D. Brooks, and T. Hernandez-Boussard, "Architecture and implementation of a clinical research data warehouse for prostate cancer," *eGEMs*, vol. 6, no. 1, 2018.

[11] S. K. Sahu, A. Anand, K. Oruganty, and M. Gattu, "Relation extraction from clinical texts using domain invariant convolutional neural network," *arXiv preprint arXiv:1606.09370*, 2016.

[12] S. Sohn, J.-P. A. Kocher, C. G. Chute, and G. K. Savova, "Drug side effect extraction from clinical narratives of psychiatry and psychology patients," *Journal of the American Medical Informatics Association*, vol. 18, no. Supplement_1, pp. i144–i149, 2011.

[13] C. Quan, M. Wang, and F. Ren, "An unsupervised text mining method for relation extraction from biomedical literature," *PloS one*, vol. 9, no. 7, p. e102039, 2014.

[14] N. Kang, B. Singh, C. Bui, Z. Afzal, E. M. van Mulligen, and J. A. Kors, "Knowledge-based extraction of adverse drug events from biomedical text," *BMC bioinformatics*, vol. 15, no. 1, p. 64, 2014.

[15] L. R. Pillai, G. Veena, and D. Gupta, "A combined approach using semantic role labelling and word sense disambiguation for question generation and answer extraction," in *2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC)*. IEEE, 2018, pp. 1–6.

[16] G. Veena, D. Gupta, A. N. Daniel, and S. Roshny, "A learning method for coreference resolution using semantic role labeling features," in *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2017, pp. 67–72.

[17] R. A. Das, P. Afsal, and M. Thushara, "Tagging of research publications based on author and year extraction," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2018, pp. 892–896.

[18] V. G. Lekshmy, P. Anusree, and V. Varunika, "An implementation of genetic algorithm for clustering help desk data for service automation," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2018, pp. 952–956.