

Exercise 2 : Literature Review

Relation Extraction in Clinical Text using NLP Based Regular Expressions

Sunil Gauda
ID: 10595858

Assessment


Clinical Data comprises keywords which define the various aspects of the domain. A keyword can be a name of the disease, medication, treatment, symptom and many other things. These keywords usually connect together to define a diagnosis of specific illness and treatment of it. Using these keywords, Natural Language Processing and programming techniques Author Veena G, Hemanth R and Jatin Hareesh, developed a method of creating relationships between the keywords, which will help identifying, simplify the various processes of the healthcare domain.

The key part of this research is data, in this paper the author has acquired the data using python based web scraping on websites related to medicine, and Word-Net Lexical Database which provides with the definition and meaning of the words extracted, Authors have used python and machine learning to derive meaning from clinical texts and set up a pipeline which uses various techniques to extract, process, Tag, Label, and derive mathematical similarity of data. The process uses various techniques on each stage to provide an appropriate process of relationship definition and extraction between nouns, which can be singular or plural in nature.

Web Scraping is done using python, it involves scraping websites to extract relevant data for the process, which includes websites with general information about clinical data like Wikipedia or other focused websites for clinical data, then Wordnet data is scraped to get the meaning out of the extracted text using syn-sets, The extracted data is basically a text document and the text which is used for further process.

Regular expressions are program syntax which we can use to modify or extract information from text, python as a module called “re” which the Authors have used to clean the text data, Author basically converted the text data to keyword based related data, which completed the process of having structured data for tagging.

The data/corpus we have now can be used to identify the words by parts of speech, is used, Part of Speech (POS) Tagging which helps Author's identify nouns, pronouns, verb, adjectives and various parts of speech, helps Author's clear our data of non related data in terms of clinical keywords. This thereby completes the process of data augmentation, hence the resultant outcome is concise data which can be labeled and processed on.



Labeling is done to provide unique signatures to each word, this is the part of NLP where the keywords are labeled to create groups of related data, in this case organs are part of body, Author's have used "Meta-Map" library to label the keywords that are clinical in nature, this helped them grouping all the related words together, like organ-body, bleeding-symptom, the process until now has created a processed data which is meaningful and can be used to create relationship between them, this is where authors have concluded data processing.

Path-Similarity is used to redrive relationship between words, but to derive relationship it is required to have syn-nets, and hierarchy of it, using Word-net, authors have drawn final product of the relationship extraction which is defining relationship between the words of the preprocessed data, now with the hierarchy of data sorted clear sequence of the clinical data is acquired like symptoms can be sweating, bleeding, fatigue.

The paper is based on previous work of Veena G et al., have derived work from A. Makowiecka, M. Marciniak and A. Kupsc in regards to the rule based data retrieval system using clinical records and mammogram reports with understanding of grammatical composition in medical documents. They improved upon it with the techniques of concept extraction by X. Fu and S. Ananiadou which uses machine learning for entity identification in the records, which was mainly discharge summaries and progressive notes of the patient in the hospital with pre/post - operating procedures, similar work was also done by W. T. Abdel-moneim, M. H. Abdel-Aziz, and M. M. Hassan on patient narratives using clinical texts. Further modifying the techniques of information extraction D. Demner-Fushman and J. Lin used question and answer techniques to derive information for relationship management in the domain of proof based medicine, The last work to which this paper refers to is B. Rink, S. Harabagiu, and K. Roberts who have used a supervised machine learning system for similar relationship extraction task on medical records.

Conclusion

Author Veena G et al., have derived the relationships using clinical data scraped from websites, and have provided an example of working hierarchy and relationship identification of the data which can help identify and diagnose illness faster and help reduce health crises, but the definition of the infrastructure of the process is not completely added to the paper, and the paper lacks a mathematical explanation of POS and PS procedures used in the process.