CHAPTER -1

INTRODUCTION

1. 1 OVERVIEW

The automatic extraction task of structured information from the semi structured and unstructured documents is known as Information Extraction (IE). The main goal of Information Extraction Process is to extract the predefined set of concepts in a particular domain. It will ignore the irrelevant information. The domain consists of set of documents which is known as a corpus, along with the specifications of needed information. Deriving structural information from the unstructured is known as Information Extraction.

The Information Extraction is a text mining process which consists of Information Extraction and Information Retrieval. The Information Retrieval process is to retrieve textual documents from a collection of documents[8]. The Information Retrieval process retrieve a subset of documents based on the query submitted by the end user. The query will be generated based on the keywords which are related to the specific need of the end user.

On deploying Information Extraction on text it converts unstructured form of information into structured form of information. The Information Extraction process includes various subtasks such as:
1.2 INFORMATION EXTRACTION VS INFORMATION RETRIEVAL

Information Extraction is not similar to Information Retrieval. Information Retrieval is the process of identifying documents relevant to the requirements specified by the users. The subset of documents related to the users’ requirements will be produced by the Information Retrieval System as output.

Fig 1.1: Information Retrieval

The Fig 1.1 illustrates the Information Retrieval whereas Fig 1.2 illustrates the Information Extraction Process also shows the difference between Information Retrieval and Information Extraction.

The Information Extraction will be the toughest job than the Information Retrieval. The Information Retrieval required only the requirements from the user. The Information Extraction need to specify
in detail about the features and its various classifications of extracted features and the relation between the identified entities[10].

Fig 1.2 : Information Extraction

Throughout the last decade, biomedicine has expanded at an incredible speed. Every day a remarkable number of biomedical documents were published and a huge quantity of novel information was created. The Natural Language Processing (NLP) systems required to process the huge amount of information generated by the biomedical domain. The NLP tools are utilized to process the biomedical literature to identify the various interactions between the protein and the genes and also to analyze the Structure of protein and its functionalities, the relation between the protein and the genes[78][9].

The most important content of information from the biomedical literature is the newly identified gene types [1][78], protein types and the new relations and interactions as the outcome of new research.
Enormous amount of research is going on to discover the gene, protein and role of DNA sequence. The outcome of these researches were published in the biomedical literature. The knowledge hidden in the biomedical literature is more valuable for the researchers, biologist and physicians. Very few tools are available to represent and visualize the information available in the biomedical literature. Due to the huge volume of information available, a reliable automatic system to extract and process the information is desirable. This kind of automated system updates the information-base. The purpose of the research work is to generate an automated system to extract the named entities and relations between the entities from the biomedical documents[76].

1.3 TEXT MINING PROCESS

Text Mining is a Data Mining Process, which is used to obtain high quality information from the text. The text mining analyzes the document and identifies the structured information's from the set of information's. The Text mining consists of various subtasks namely

- Information Retrieval
- Information Extraction

The Information Retrieval process will retrieve sub set of documents from the set of documents based on the user need., whereas Information Extraction extracts structured information's from the subset of documents.
The documents satisfying the user requirements will be retrieved from the information-base. From the retrieved set of documents, a corpus will be generated.

**Fig 1.3: Information Retrieval and Information Extraction**

The documents available in the corpus undergoes various processes in the preprocessing phase. After preprocessing using Natural Language Processing tools and various machine learning algorithms [40][72], the information will be extracted from the preprocessed documents. The Fig 1.4 : illustrates the typical Text Mining process.

**Fig 1.4: Typical Text Mining Process**
The Typical Text Mining process consists of various subtasks such as:

- Data Acquisition
- Text Preprocessing
- Modeling
- Evaluation
- Application

Text Mining consists of retrieval of information and extraction of information. The Information Retrieval is the task of retrieving the set of documents based on user requirements and generates the corpus. Information extraction process consists of various subtasks such as:

- Named Entity Recognition
- Co-reference Resolutions
- Relation Extraction

1.4 NAMED ENTITY RECOGNITION

Named Entity Recognition is a technique for identifying entities that is proper nouns from the given text and classifying them based on the different categories. Common kind of Named Entity categories are person names, locations, organizations, designations, address, etc.. Named Entity Recognition is not a simple process. It is a sequence of various sub processes. Starting from identifying the sentences and analyzing the patterns of various sentences. understanding various
parts of speech of sentences. Recognizing the named entities of various
types and classifying them based on various categories.

The Named Entity Recognition process involves various methods
such as Dictionary based approach, Rule based approach, Pattern
recognition approach, Machine Learning based approach etc... And
various Natural Language Processing Tools are also included in the
process of entity identification process [17]. Named Entity Recognition
is not a single process, it is a sequence of various subtasks.

1.5 RELATION EXTRACTION

The Natural Language Processing tools include tokeniser,
Sentence splitter, Parts-Of-Speech tagger, Noun Phrase Chunker etc..
To extract relations from the biomedical text various challenges that
arises[10]. Due to high ambiguity of in vocabularies, sentences which
are too long and complex. Biomedical entities and relations differ from
other general terms[14]. The NLP tools normally trained on general
English text than the biomedical text. Fig 1.5 exemplifies the relation
extraction system.

Another major drawback is the lack of gene and protein name
standardization and huge possibility of diversity in biomedical
terminologies.
Fig : 1. 5 A typical Relation Extraction System

This will have an impact on the recognition of named entities from the biomedical text. Because of these challenges, extracting relations from biomedical text has been an active research field in the last decade[9].
The process of analyzing the Natural Language text is known as Natural Language Processing, involved in converting unstructured text into structured text. Natural Language Processing is consists of

- Lexical Analysis
- Syntactic Analysis
- Semantic Analysis

The first phase lexical analysis converts unstructured text into various tokens. Using the morphological structure the word form is identified in the lexical analysis. The Syntactic analysis phase adds the parts-of -speech tags to the text. In semantic analysis, the Natural Language Processing techniques were included to annotate the text.

Though various methods have been proposed to recognize biomedical entities and biomedical relations from the biomedical documents still remains a big issue due to the excellence of the recognized relations, time taken to recognize and the kind of relations being extracted[27][37]. Initially, the performance of extraction systems is considered in terms of precision, recall and F-measure. Earlier systems focused on identifying only the interactions between Proteins, but there are much hidden information still untouched in the field of biomedical research.
### 1.6 OUTLINE OF THE THESIS

The remaining part of this thesis is organized as stated below.

**Chapter 2** provides literature survey which is the background for entity and relation extraction methods in biomedical documents, comprises of a detailed literature review of recent works related to the proposed work.

**Chapter 3** presents the preliminary concepts of the proposed system. Starts with an overview of a typical entity and relation extraction system and its overall proposed architectures such as its workflow. Tools and techniques which are used in various processes.

**Chapter 4** deals with the Information Retrieval from the biomedical database in the unstructured format based on the user requirement specifications. This presents a novel method to extract biomedical documents from the biomedical database and converts the structure of the documents into the format which can be understood by the Natural Language Processing tools and generate a corpus of biomedical documents.

**Chapter 5** introduces a new method to extract the named entities from the biomedical documents. The new method is a hybrid method which combines rule-based and machine learning based approaches.[78] This
new method uses a set of features with Conditional Random Field algorithm.

**Chapter 6** a novel method to extract the relations from the biomedical documents is discussed in this chapter. This method uses the co-occurrence and machine learning methods to extract the known and unknown relations from the biomedical documents.

**Chapter 7** Finally, discusses and evaluates the outcomes of named entity recognition and relation extraction based on the requirements specification opted by the users. The results were compared with various existing methods.

**Chapter 8** Finally concluded the work presented in the thesis and also conversed the road for possible future work.