CHAPTER -8
CONCLUSION & FUTURE ENHANCEMENT

8.1 OVERVIEW

The extraction of biomedical entities and relations from the biomedical documents is a well studied problem. Different approaches for entity and relation extraction have been proposed[11][77][73][16]. The proposed system for named entity recognition and relation extraction is framework that consists of various sequence processes[73]. The proposed system consists of subsystems such as Information Retrieval (IR) and Information Extraction (IE). The IR and IE process incorporate various subtasks.

8.2 INFORMATION RETRIEVAL

The goal of IR system is to retrieve the documents from the biomedical databases such as PubMed, MedLine etc,.. belong to the NCBI which is known as National Center for Biotechnology Information belonging to United States National Library of Medicine[77]. The IR system gets the user requirements and makes the query. Using the query, documents are retrieved and stored in the corpus. The retrieved corpus is structured and validated with relevance to Document – Weight Matrix. The process of validating the relevance is used to avoid irrelevant and unwanted documents in the corpus. This helps in improving the outcome.
8.3 INFORMATION EXTRACTION

The information Extraction process includes various subtasks to identify the expected known and unknown named entities and relations. The existing approaches for Named Entity recognition can be categorized in to three classes such as

- Dictionary based approach
- Rule based approach
- Machine Learning based approach

All these approaches having their own drawbacks. Dictionary based approach can only identify the known entities which are available in the existing dictionaries using the pattern matching technique[74]. The growth in biomedical research day by day leads to engender more components and their associations.

Rule-based approach also has its own limitations. The rule sets are crafted manually. This will work properly for minimal set of corpus and also it is difficult to identify the unknown entities. Machine Learning approach works well for biomedical extractions but it needs huge set of training data[76][6].

The proposed approach is a hybrid approach. Which is the combination of rule-based approach and machine Learning based approach. The hybrid approach overrule the drawbacks of the existing systems.
The feature space based Conditional Random Field machine learning algorithms is used in hybrid system. In the domain of biomedical, the Conditional Random Field machine learning algorithm works properly than the other machine learning algorithms like Support Vector Machines (SVM), Hidden Markov Models (HMM), Maximum Entropy Markov Models (MEMM)[36].

Conditional Random Field machine learning algorithm has several advantages over other methods[18]. Conditional Random Field machine learning algorithms avoid the label bias problem which is a weakness of Maximum Entropy Markov Model (MEMM). The conditional nature of Hidden Markov Model (HMM) results in relaxation of the independence assumptions that can be overruled by Conditional Random Field machine learning algorithm. Support Vector Machines (SVM) requires more time to train complex models than Conditional Random Field machine learning algorithm[67][69][10].

These contributions are the crucial components to a successful Named Entity System for Biomedical documents. The results of Feature Space based Conditional Random Field Named Entity Recognition system is compared with the results of Support Vector Machines and found that the Proposed system out performs the existing systems.
8.4 Relation Extraction

The Relation Extraction system works in two phases. Initially, the co-occurrence technique is applied on the data. Then the Conditional Random Field machine learning algorithm based approach is applied[11]. The conditional Random Field is used to identify all possible relations from the biomedical corpus.

The proposed algorithm for relation algorithm is compared with the rule-based approach[47]. The proposed algorithm outperforms the other approaches for relation extraction. Though the proposed system outperforms the existing systems in many scenarios the system has its own limitations.

8.5 Future Enhancement

The proposed work to extract the entity and relation from biomedical documents using the machine learning technique Feature space based Conditional Random Field works well within its limitations. The outcome of the work compared with the machine learning algorithm Support Vector Machine. In many cases of identifying the entities of various classes, FS-CRF (Feature Space based Conditional Random Field) exhibits better results than Support Vector Machine and also the other techniques used for the same purpose.
The future work is to focus on the extended work in the mining of documents in regional languages like Tamil, Telegu, etc, to find the unknown knowledge on various heritage based medicines and medical treatments hidden in the various forms of ancient literatures.