Chapter 1

INTRODUCTION

“One machine can do the work of fifty ordinary men. No machine can do the work of one extraordinary man.” - Elbert Hubbard

1.0 INTRODUCTION

Expert systems have a wide application in various fields. One such field is medicinal field. Cancer, a deadly disease could be easily diagnosed with the help of this expert system. An expert system is used to detect cancer caused in various parts of human body especially stomach and prostate. A large number of health organizations are making use of fuzzy expert systems in detecting stomach cancer and prostate cancer [Maria, 2007].

Expert system is one of the approaches of Artificial Intelligence [Webref, 1]. Expert system is nothing but software with expertise or specialization in a particular field. It provides answers and solutions to people in making decisions under uncertain conditions.

This research work deals with AI and Natural Language Understanding for the appliance of microbiological literature.
Natural Language Understanding is one of the old research areas and the usage of this natural language understanding to molecular biology is very new field. The field of natural language understanding is robustly associated with Artificial Intelligence (AI).

Everyday more than three thousand new articles are being published and indexed. Hence biologists are necessitating to know more about existing knowledge. Right now natural language understandings in biology results are not satisfactory. Thus more concentration on research is required to share the challenges of creating useful natural language understanding application for the Biologists.

There are many approaches to information extraction from biomedical texts. One of them is the rule pedestal and grammatical which is called Natural Language Understanding (NLU). The other approach is pattern matching, known as Natural Language Processing (NLP). The goal of this project is to do full parsing i.e., mining of microbiological texts. Full parsing used to attain the goal of automatic information mining in the field of medical domain. Full parsing means each and every sentence must be wholly analyzed from the beginning to the end.

1.1 TASK SPECIFICATION

The genuine task specification of this research project is that the existing facts about genes and their prologue are available only as a text. Searching and cross linking of such information is through existing databases or indexes created manually. Hence an expert system is needed and that is able to identify the occurrences of a gene in a text correctly.

Thus the aim of this research project is to find out whether full parsing works in biomedical domain or not. It is supposed to be shown that full parsing works are better than statistical approach used in PubGene [Arzucan Ozgur, 2010]. In this project it has been shown that the quality of the extracted relations among genes [Hyun Seok Park, 2005] will be better if the extraction is based on full parsing.
In this project it focused on subset of MEDLINE abstracts a product of NCBI, which contains the word “gastrin”. It has been able to improve the values of Recall and Precision and F-measure values.

1.2 RESEARCH HYPOTHESIS AND OUTCOMES

The main research questions recognized and explored by this research work are whether computer program like Gene Understanding Machine is able to understand text about molecular biology like humans can or not.

The above research question could be breakdown as “could full parsing be used to extract information from these molecular biology texts”. Search technologies like Google could be used for this work or not.

Is the above method or approach is possible? if possible then “is this method is useful at how much percentage of the sentence could be understood by a computer and how fast the computer could understand”. If all the above hypothesis questions are possible then this approach is better than others.

1.3 STRUCTURE OF THE THESIS

The thesis is alienated in to ten chapters. The initial chapter provides introduction of the thesis. In this chapter specification of the task and research hypothesis is presented. It is also presented how the thesis is organized in the remaining chapters in the form of structure of the thesis.

The second chapter provides natural language understanding amend articles. In this chapter it enlightens the relevant linguistic terminology used for this thesis work and explained the important concepts of computational linguistic terms like ontology along with some of the constructs and problems that were encountered during the work with text from the MEDLINE databases. Some of the important components of an expert system are also presented.
Third chapter gives an idea of Research Methodology that summarizes the research contributions and the motivation. In this chapter it has shown the important snapshots and milestones of the research work towards the implementation of gene machine. It has also shown the usage of local grammars, machine erudition and finally the description application of Google application programming Interface for the verification and development of the system.

Chapter four gives information mining from Biomedical Literature initiates an impression of Natural Language Understanding [B.V.Subba Rao and K.V.Sambasiva Rao, 2009A] and shows how the information extraction crisis could be solved using local grammars. The editorials are about natural language understating and its applicability to microbiological text. In this chapter gene and protein activations are converse along with nearby research areas in the area of natural language understating. Local grammars are developed using a visualization system named Unitex, it generated some good results and that results are compared with full parsing approach.

Chapter five shows the usage of machine learning to automatically recognize protein name based on feature from the gene search engine [B.V.Subba Rao and K.V.Sambasiva Rao, 2011]. The method used in this chapter is through using an existing Medline databases for automatic information extraction from biomedical corpora. It is able to produce positive and negative instances to these databases. Protein related words are known as negative examples and protein synonyms are known as positive examples [Hong Yu, 2003]. It produces the results on Medline abstracts [Hong Yu and Vasileios Hatzivassiloglou, 2003]. Finally this chapter presents a method for creating positive and negative data for protein names.

The sixth chapter discusses an online implementation of the system and shows that the methods [B.V.Subba Rao and K.V.Sambasiva Rao, 2009C] scale well to a larger set of entities.
It also presents how the Internet search engines like Google, Yahoo, Askme and Bing used in biomedical domain. It has been observed that, how Google search engine could be used to find biomedical relevant information as protein interactions from the internet. The results presented in this chapter are extensions of the earlier works on semantic annotation of proteins. Here it recapitulates the results together with an approximation of the success rate of the system. The performance of the system is computed as Precision, Recall and F-measure values.

In chapter seven, Rationalization of Protein contact using Google and Gene Ontology [B.V.Subba Rao and K.V.Sambasiva Rao, 2010B], it presented a new idea ProgG using google application programming interface to discover annotations for some particular proteins. Based on these experiments it got good results that are 59% terms are valid for Gene Ontology annotations and 18% protein names in the solutions given by ProgG [B.V.Subba Rao and K.V.Sambasiva Rao 2010B].

Chapter eight, online mining and explanation of biomedical context using Google [B.V.Subba Rao and K.V.Sambasiva Rao, 2010A], it depicts an online execution of the system. This chapter describes an open source software tool for text mining in molecular biology tasks. It has been observed Google search engine, and how it locates biological related information on protein associates from the net.

In chapter nine, Language Sympathetic in Molecular Biology expresses the system that uses natural language processing in order to study the abstracts of the researcher paper automatically [B.V.Subba Rao and K.V.Sambasiva Rao, 2009B]. In this chapter it has been shown that how this system came into existence to handle large collection of biomedical abstracts and able to shows good performance in retrieving the required gene related abstract information to biomedical researchers.
As a final point it has been shaped an expect system Gene Understating Machine (GUM) that could support biologists and other research scholars in dealing with the ever escalating information overload in their filed.

This research project proved that both Google API and Gene Understanding Machine are important systems that could play a role in fulfilling the dream of real automatic information extraction.

The final chapter gives the fine points of results, discussion, conclusions and future scope of work.

1.4 SUMMARY AND CONCLUSIONS

To achieve the above task specification a system is initiated i.e., An Expert system based Gene Understating Machine for Natural Language Applications and slowly improved the system step by step and finally it reached the goal. The step by step improvements of the system is shown in the thesis outline as a collection of ten chapters.

1.5 PUBLICATIONS

➢ In International Journals

1. B.V.Subba Rao and Dr. K.V.Sambasiva Rao “Intuitive Information Mining from Biomedical Literature” a research paper published in International Journal on Computer Engineering and Information Technology (IJCEIT), Volume 6, No 10, Pages 30-35, 2009A.

2. B.V.Subba Rao and Dr. K.V.Sambasiva Rao “Mining and Elucidation of Biomedical Context Using Google” a research paper published in International Journal of ETA-ETS, Volume 3, Issue 1, Pages 358-361, Jan 2010A.

- In National Journals

1. B.V. Subba Rao and Dr. K.V. Sambasiva Rao “The GUM and Google: Natural Language Understanding in Molecular Biology Text” a research paper published in ANU Journal of Engineering and Technology, Volume 1, Number 1, Pages 17-21, June 2009.


- In International Conferences


- In National Conferences