Design and Development of Automated Ontology from PubMed abstracts using Rule based approach

G. Suganya
Department of Computer Science
Bharathiar University
Coimbatore, India
suganyadheksha@gmail.com

R. Porkodi
Department of Computer Science
Bharathiar University
Coimbatore, India
porkodi_r76@buc.edu.in

***Abstract*: Ontology is an emerging discipline that has the huge potential to improve information in organization, management and understanding. It has a crucial role to play in the field of information extraction and information retrieval. Gene names extraction is an important problem in the area of biomedical field through which the hidden associations among genes, diseases, mutations and drugs can be extracted and helpful in solving many diseases in human. This paper developed a framework to build an automated ontology from the PubMed abstracts. The design and development of automated ontology consists of two important phases: Identifying the gene names using set of rules and construction of automated ontology from the identified gene names. This work uses 100 PubMed abstracts randomly. The developed automated ontology extracted and visualized the gene names using DLquery and the gene names extracted by this framework using rule based approach compared with existing Genia tagger.**

Keywords: Ontology, Genia tagger, PubMed abstracts, Gene, Regular expression

1. INTRODUCTION

Ontology is a controlled vocabulary of well-defined terms with specified relationships between them capable of interpretation by both humans and computers. It is used to recognize the specified entities relationships between them and also used to describes the structure of the information and defines an ordinary vocabulary to share information in a domain. The reason for preferring ontology is to provide a formal specification of biomedical knowledge, classification of biomedical entities, develop a common understanding of the entities for a known domain, reuse of data and knowledge, explores the domain knowledge and different domain knowledge from the operational information, easy to identify and update legacy data and also distribute the information. The ontology can be created in various tools such as Protégé, Text2Onoto, IsaViz, Apollo and SWOOP. The ontology contains the information about the relationships between concepts and used to express semantic similarities.

The main objective of the work is to develop an automated ontology using rule based approach from PubMed abstracts. The paper is organized as follows: section 2 discusses the related work, section 3 focuses on the framework of automated ontology creation, section 4 discuss the results and discussion and finally the work is concluded in section 5.

2. LITERATURE SURVEY

KamelNebhi (KamelNebhi 2012) presented an ontology-based information extraction system from twitter using rule-based approach. For the ontology-based information extraction system, balanced distance metric is used to analysis the performance measure. Xinhou et al (XinHou, S.K. Ong, A.Y.C. Nee, X.T. Zhang, W.J. Liu 2011) proposed an automated domain ontology algorithm from the domain corpus named as “GRONTO”. The created ontology was compared with the FIXON system. The proposed system obtains the more number of precision and recall value than the existing system. Dai Quoc Nguyen et al (Dai quoc Nguyen, DatQuocNgyen, KhoiTrong Ma, Son Bao Pham 2012) developed a system that automatically builds ontology from Vietnamese texts using cascades of annotation based grammars. The created ontology was viewed by using graph visualization named as OntoGraf with plugin tool. Ontology checked the consistency by using reasoner and retrieves the query by using DL Query.

Ayesha Ameen et al (Ayesha Ameen, KhaleelUr Rahman Khan, B.Padmaia Rani 2012) created ontology for university. The university ontology was checked the consistency by using the reasoner and various restrictions are applied to classes. The created ontology can be integrated into any university to facilitate efficient access and retrieval of information that can be processed automatically. Koning D and Sarkar (Koning D, Sarkar I, Moritz T)has proposed tool which consists of various rules based on the regular expression. It identifies the all words that are not in the common dictionary and it also find whether the term is a species or not. Yu H (Yu H, Hatzivassiloglou V, Friedman C, Rzhetsky A, Wilbur WJ) et al proposed the method to retrieve synonyms of proteins and genes from abstracts as well as full text and identified more number of synonyms with high accuracy.

Hanisch D, Fundel K (Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J) et al established a technique to find the protein and gene names by using rule based tool named as "Prominer". It includes the various spelling variants to support gene name matching including an approximate matching string as a sequence of tokens which are assigned to corresponding classes. The classes are used to find the mismatches in the appropriate matching.Hong Y(Hong Yu, a, Vasileios Hatzivassiloglou, Andrey Rzhetsky, W. John Wilburc) et al founded the synonyms of the identified gene, protein names from the abstracts. The gene and protein names are usually represented by symbols. The names usually are in the long forms of their symbols and describe the functions of the gene/ protein. R Porkodi and B L Shivakumar(Hong Yu, a, Vasileios Hatzivassiloglou, Andrey Rzhetsky, W. John Wilburc) have proposed an approach is to construct gene and protein names dictionary using rule based approach from Medline abstracts. The regular expressions are generated for finding the gene and protein names and also developed a set of pattern matching rules that relates to the gene and protein symbols to the corresponding names. This achieves the 81% accuracy in identifying the gene and protein names.

3. FRAMEWORK OF CREATING AUTOMATED ONTOLOGY

Dataset description: The input dataset is downloaded from National Centre for Biotechnology Information (NCBI) website which contains the PubMed abstracts and the set of PubMed abstracts in which the gene names are extracted using the regular expression. For the experimental purpose 100abstracts are used. The framework of automated ontology is shown in Figure1.

The framework consists of four phases namely Pre-processing, identify and extract gene names from the PubMed abstracts and creating automated ontology creation by concept validator and concept adder tasks. The final phase concentrates on the visualization of automated ontology and the extracted gene names by rule-based approach compared with Genia tagger.

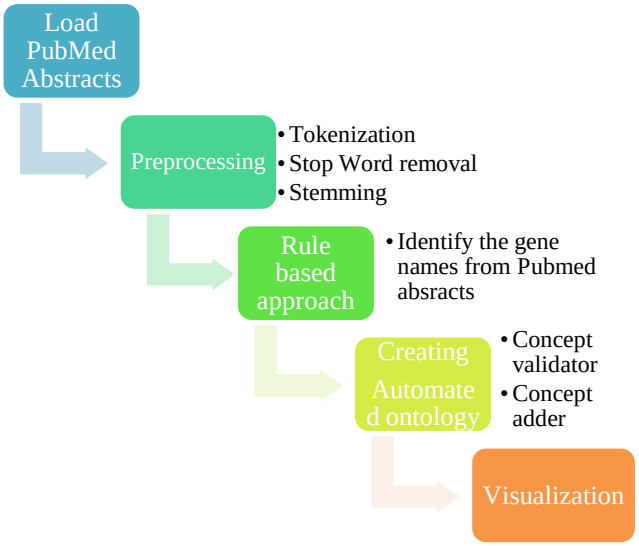

Figure 1. Methodology for creating automated ontology

3.1 Pre-processing

Three pro-processing techniques are applied to PubMed abstracts such as tokenization, stop word removal and stemming.

3.1.1 Tokenization

Tokenization is the process of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and etc. In this process, the punctuation marks are removed.

3.1.2 Stop word removal

The stop words refer to the common words in a language like “and, are, this, etc”. They are not useful in classification of documents. So they have to removes from the text document. There is no single universal list of stop words are used in NLP tools. This process reduces the text data and improves the system performance.

3.1.3 Stemming

Stemming is the process of conflating the variant forms of a word into a common representation. For example the words “presentation, presented, presenting” could all be reduces to a common representation “present”. This technique is widely used to produce the text processing for information retrieval.

3.2 Identifying the gene names using Rule based Approach

The rules are constructed using regular expression that are used to extract the gene names from the PubMed abstracts. There are 14 rules are framed using regular expressions and showed in Table 1.

Table 1: Regular expression for extracting gene names

S.No	Rules	Regular Expression	Example
1	The word of the letter matches the full capital letter of each word in the full name	\b[A-Z]{2,9}\w*\b	BRAF
2	The word of the letter matches the full capital letter followed by one or more numbers	[A-Z]\w*[0-9]	GRP78
3	The word of the letter matches the full capital letter followed by numbers and full capital letters	(A-Z]\w*\d)-([A-Z]\w*)	FLT3-ITD
4	The word of the letter matches the first letter capital and followed by numbers and last letter is capital letter	[A-Z]\d*[A-Z]\w*	F691L
5	The word of the letter matches the first letter capital and followed by numbers	([A-Z])-(\d*)	G-749
6	The word of the letter matches the first letter is capital letter and followed by lower case and letter is Arabic numeral	([a-zA-Z])+(\d+) or [A-Z][a-z]\w*[0-9] | [a-z]\w*[0-9]	Rac1
7	The word of the letter matches the two words with special characters; first word all letters are capital and second word full capital followed by Arabic numerals	[A-Z]\w*-[A-Z]\w*[0-9]	SHH-GL1
8	The word of the letter matches the first letter is capital and followed by lower case	[A-Z]\w*[a-z]	Pak
9	The word of the letter matches the first letter is capital and followed by lower case letters and last letter must be capital	[A-Z]\w*[a-z][A-Z]	RhoA
10	The word of the letter matches the first word of first two or three letters are capital and followed by lower case and second word have the Arabic numbers with special character	[a-zA-Z]\w*[a-z]-\d	EHop-016
11	The word of the letter matches the first two letters are capital and followed by number and last letter is capital	[a-zA-Z]\w*\d[A-Z]	PI3K

S.No	Rules	Regular Expression	Example
12	The word of the letter matches the first word and second word of the first letter is capital and followed by lower case with special character	[A-Z]\w*[a-z]-[A-Z]\w*[a-z]	Bcr-Abl
13	The word of the letter matches the two words with special character; first word full capital letters and second word full capital letters followed by Arabic numerals	\b[A-Z]{2,9}\w*\b-[A-Z]\w*[0-9]	NVP-LDE225
14	The word of the letter matches the first letter is lower case and followed by upper case letters	[a-z]+[A-Z]\w*	mTOR

The rule 1 is used to extract the gene name where all letters of a word in the capital. The gene names are generated using the rule-based approach. The rule 2 is used to find the word in the form of starting with capital letter and ending with Arabic numerals. The rule 14 is used to extract the gene name which will be in the form of first letter is lower case and followed by all upper case letters. The gens extracted from PubMed abstracts using these 14 rules in rule based approach compared with Genie tagger.

3.3 Automated ontology creation

The creation of automated ontology consists of two main tasks: Concept validator and concept adder. The concept validator will check whether the concept or gene extracted from PubMed abstracts is already existing in the ontology or not. If, there is no need to add; otherwise add the concept or gene into the ontology. The concept validator must be involved before adding the concepts into the ontology.

The automated ontology was done with the following steps:

Input: *Ontology X and extracted terms from PubMed abstracts*
Output: *Updated automated ontology X*
Steps:
Let X be automated ontology with concepts $X_1, X_2, X_3, ...X_n$
Let Y be the extracted gene names $Y_1, Y_2, Y_3, ...Y_n$ from PubMed abstracts
For each Y_i in Y
Do
//Concept validator
For i=1 to n
Begin
If $X_i <> Y_i$ then
//Concept adder
Add the concepts X_i into the Ontology Y
Else
Break;
End
End

3.4 Visualization and Evaluation

Visualization is another phase that is used to visualize the concepts such as gene names in the created ontology. The extracted terms are compared with existing Genia Tagger.

4. RESULTS AND DISCUSSION

The identification and extraction of terms from PubMed abstracts are evaluated based on accuracy measure. The accuracy is calculated based on the number of terms identified correctly from PubMed abstracts. Totally 14 regular expression are used to identify and extract the gene names. The figure 2 shows the four PubMed abstracts of size 5MB, 10MB, 15MB and 20MB respectively and terms the extracted by the Rule based approach and Genia tagger. It is found that the proposed rule based approach works extract more than 50% of gene names as same as Genia tagger. The accuracy can be improved by adding appropriate rules which are used to recognise all variant representation of gene in the PubMed abstracts.

Table 1. Extracted terms count

PubMed abstracts size	Extracted Genes	
	Rule based approach	Genia tagger
5 MB	12	21
10 MB	15	32
15 MB	20	40
20 MB	30	55

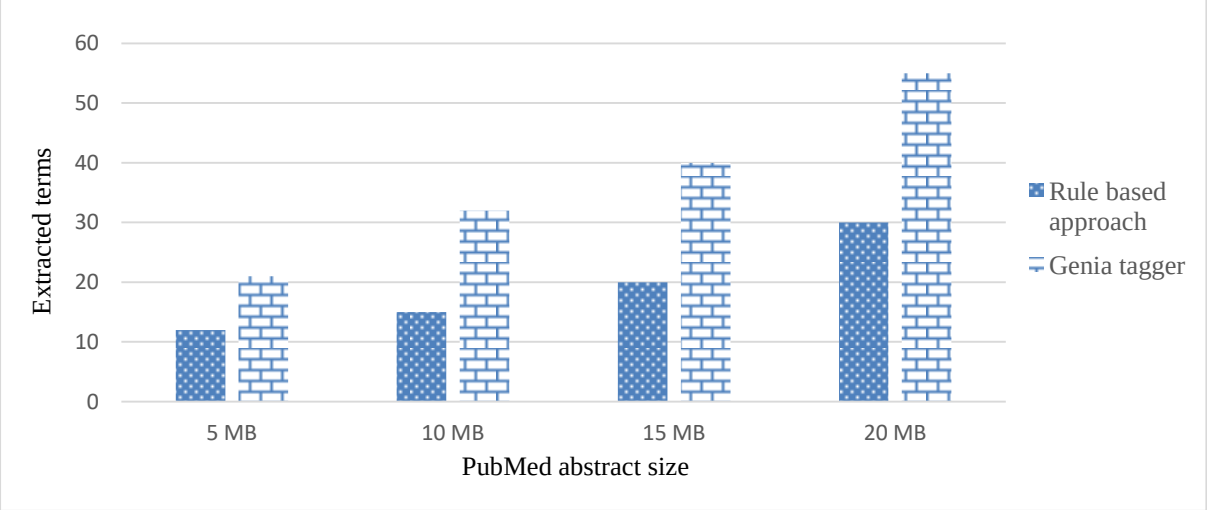

Figure 2. Comparison with the Genia tagger

```
<?xml version="1.0" encoding="utf-16"?>
<Ontology xmlns="http://www.w3.org/2002/07/owl#" xml:base="http://www.semanticweb.org/hp/ontologies/2017/7/untitled-ontology-58"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:xml="http://www.w3.org/XML/1998/namespace" xmlns:xsd="
http://www.w3.org/2001/XMLSchema#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" ontologyIRI="
http://www.semanticweb.org/dheksha/ontologies/2018/6/untitled-ontology-94">
<Prefix name="" IRI="http://www.semanticweb.org/dheksha/ontologies/2018/6/untitled-ontology-94" />
<Prefix name="owl" IRI="http://www.w3.org/2002/07/owl#" />
<Prefix name="rdf" IRI="http://www.w3.org/1999/02/22-rdf-syntax-ns#" />
<Prefix name="xml" IRI="http://www.w3.org/XML/1998/namespace" />
<Prefix name="xsd" IRI="http://www.w3.org/2001/XMLSchema#" />
<Prefix name="rdfs" IRI="http://www.w3.org/2000/01/rdf-schema#" />
<Declaration><Class IRI="#Alzheimer" /></Declaration>
<Declaration><Class IRI="#Parkinson" /></Declaration>
<Declaration><Class IRI="#Migraine" /></Declaration>
<Declaration><Class IRI="#Multiplesclerosis" /></Declaration>
<Declaration><Class IRI="#Epilepsy" /></Declaration>
<Declaration><Class IRI="#APP" /></Declaration>
<SubClassOf><Class IRI="#APP" /><Class IRI="#Alzheimer" /></SubClassOf>
<Declaration><Class IRI="#APP" /></Declaration>
<SubClassOf><Class IRI="#APP" /><Class IRI="#Multiplesclerosis" /></SubClassOf>
<Declaration><Class IRI="#APP" /></Declaration>
<SubClassOf><Class IRI="#APP" /><Class IRI="#Migraine" /></SubClassOf>
<Declaration><Class IRI="#PARK7" /></Declaration>
<SubClassOf><Class IRI="#PARK7" /><Class IRI="#Alzheimer" /></SubClassOf>
<Declaration><Class IRI="#HMG-CoA" /></Declaration>
<Declaration><Class IRI="#p53" /></Declaration>
<Declaration><Class IRI="#FIP1L1-PDGFRα" /></Declaration>
<Declaration><Class IRI="#Vav1" /></Declaration>
<Declaration><Class IRI="#Rac1" /></Declaration>
<Declaration><Class IRI="#Rac2" /></Declaration>
<Declaration><Class IRI="#Bcr-Abl" /></Declaration>
```

Figure 3. XML file

The automated ontology developed in vb.net and stored in the file type as XML. The figure 3 represents the snapshot of the created XML file. The created ontology focusses on the 5 types of neuro disorder such as Parkinson, Alzheimer, Epilepsy, Multiple sclerosis and Migraine. If the genes are comes under any one of the neuro disorder then it added into the subclass for the corresponding neuro disorder name. Other gene names are added into the “other disease type”.

The visualization part visualizes the created automated ontology in protégé tool. The figure 4 shows the types of neuro diseases and its associated genes, etc. the created ontology has provided with set of object properties which are used to relate the subtypes with its gene names. These are very helpful in identify the hidden significant association among terms in the created ontology.

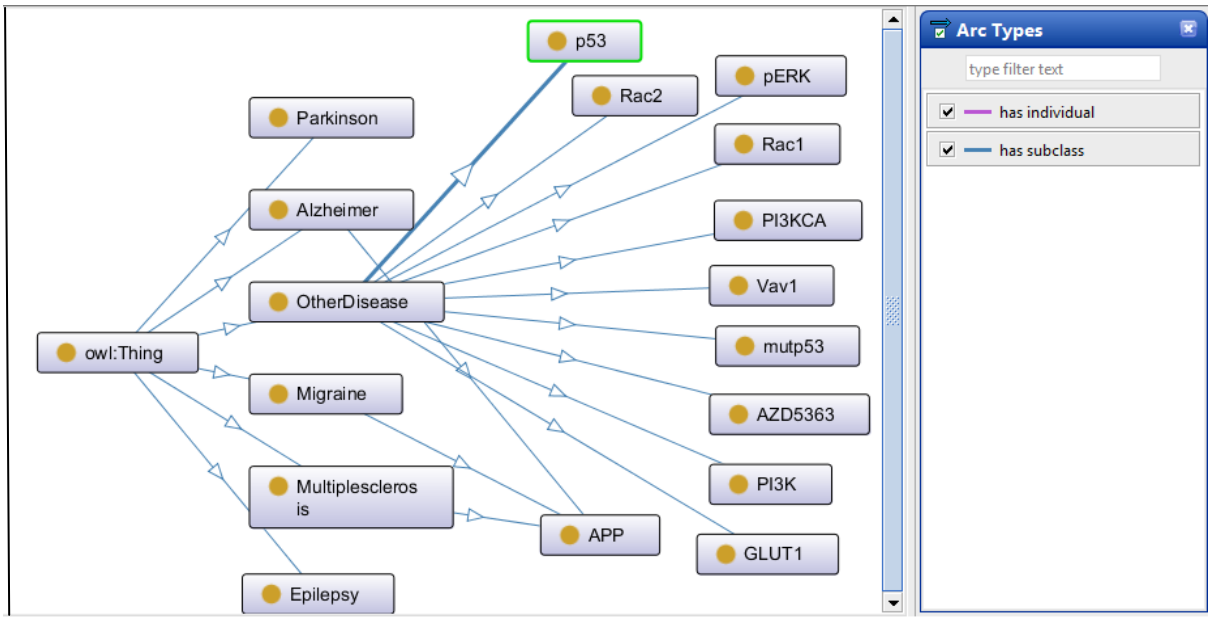

Figure 4. Visualization in Protégé

5. CONCLUSION

The hidden relationships can be easily retrieved using ontology. Most of the approaches provides only the direct relationships prevailed in the PubMed abstracts, the hidden associations cannot be extracted. Thus the ontology based informational retrieval placed a vital role. In this study, automated ontology is constructed from PubMed abstracts using rule based approach. The process involved in this research work are pre-processing, identifying and extracting the gene names, constructing the automated ontology and visualization. There are 14 regular expression are used to identify and extract the gene names. The developed approach was evaluated with the four PubMed abstracts of size 5MB, 10MB,15MB and 20MB respectively. The terms are extracted by the Rule based approach and Genia tagger. The proposed rule-based approach extracts more than 50% of gene names as same as Genia tagger.

In future, the accuracy of the extraction of gene names to be improved by apply some more rules and further this work also include drug, mutation for the respective diseases to extract better significant associations.

REFERENCES

KamelNebhi (2012), Ontology base information extraction from twitter.

XinHou, S.K. Ong, A.Y.C. Nee, X.T. Zhang, W.J. Liu (2011), GRAONTO: A graph-based approach for automatic construction of domain ontology.

Dai quoc Nguyen, DatQuocNgyen, KhoiTrong Ma, Son Bao Pham (2012), Automatic ontology construction from Vietnamese Text.

Ayesha Ameen, KhaleelUr Rahman Khan, B.Padmaia Rani (2012), Construction of university ontology

Koning D, Sarkar I, Moritz T: TaxonGrab Extracting taxonomic names from Text.

Yu H, Hatzivassiloglou V, Friedman C, Rzhetsky A, Wilbur WJ , Automatic extraction of gene and protein synonyms from MEDLINE and journal articles

Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J, ProMiner: rule-based protein and gene entity recognition

Hong Yu, a, Vasileios Hatzivassiloglou, Andrey Rzhetsky, W. John Wilburc, Automatically identifying gene/protein terms in MEDLINE abstracts

R Porkodi and B L Shivakumar, Rule based approach for constructing Gene/Proteinnames Dictionary from Medline abstract

Latifur Khan and Feng Luo, Ontology Construction for Information Selection

Wiki.csc.calpoly.edu/ontology