International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Hybridization of Max.C-Value and Permutation for Refined Concept Extraction

R. Manimala. M. C. A, M. Phil[1], Dr. G. Muthu Lakshmi. M. E. Ph. D[2]

Submitted:02/05/2022 **Revised:**20/06/2022 **Accepted:**28/06/2022

Abstract: Ontology is the representation of knowledge with a set of concepts and the relationships between those concepts, for a domain. It describes about the domain.It is mainly used to organize the web data for maximizing the semantic access and to extract the knowledge from different format of data. Intelligence is necessary for the creation and processing of those semantic metadata. Manual ontology construction is labour-intensive, error-prone process, rigid, expensive, time consuming and complex task. Ontology Acquisition or ontology learning includes the automatic extraction of domain's terms, concepts and the relationships between those concepts from the corpus of text, and encoding them with an ontology language for easy information retrieval. In the automatic ontology learning, Concept extraction is the main step to improve the accuracy .This paper describes the concept extraction using Max.CVPC method. Our proposed work is the Hybridization of Maximum C-Value of Permutation of UMLS Concepts which is further filtered with the threshold to extract the dominant concepts or refined concepts. It improves the precision as well as extracts the dominant concepts to build the appropriate ontology. Here we used Permutation technique instead of n-gram to increase the sub terms and its weightage.The method is analyzed with Genia Corpus which contains 2000 MEDLINE abstracts. The results are compared with the n-gram technique which is also filtered with maximum of C-Value. We evaluate our work with the metrics precision, Recall and F-Measure.

Keywords—*Ontology, ontology learning, Ontology Acquisition*

I.INTRODUCTION

Ontology transforms the web of documents to the web of data. Artificial intelligence and statistics help to extract concepts from artifacts. The conversion of words to concepts has been performed using a thesaurus. The mappings of words to concepts are often ambiguous. Typically each word in a given language will relate to several possible concepts. Humans use context to disambiguate the various meanings of a given piece of text, where available machine translation systems cannot easily infer context. There are many techniques for disambiguation such as linguistic analysis of the text and the use of word and concept association frequency information that may be inferred from large text corpora. Recently, techniques that based on semantic similarity between the possible concepts and the context have appeared and gained interest in the scientific community. The refined concepts are used to develop concept hierarchies, which are used to structure information into categories, thus nurture the semantic search. It formulates the rules as well as relations between the information to extract the knowledge.

Fig.1summarizes different steps required to accomplish ontology from unstructured text. The process of ontology acquisition starts by extracting terms and their synonyms from the unstructured text. After pre-processing, the related terms are extracted. Then corresponding terms and synonyms are combined to form concepts, which is the knowledge about the entity. Later the taxonomic and non-taxonomic relations between these concepts are established. Finally, axiom schemata are instantiated and general axioms are extracted to learn the rules[7]. The inference describes about the domain which is further transformed into ontology language such as OWL etc.Thus the ontology is augmented with instances of

[1]*Research Scholar(Register No.: 18124012162006), Department of Computer Science and Engineering,*

Manonmaniam Sundaranar University, Tirunelveli.

[2]*Assistant Professor, Department of Computer Science and Engineering,*

Manonmaniam Sundaranar University, Tirunelveli.

[1]*manimala22.r@gmail.com*

[2]*lakshmi_me05@yahoo.co.in*

concepts and properties. This whole process is known as ontology learning layer cake [10].

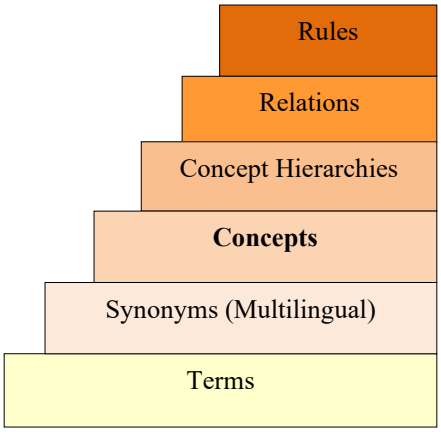

FIG.1 ONTOLOGY LEARNING LAYER CAKE [2]

This paper focuses only on Concept extraction and their refinement process which improves the ontology construction.

II. LITERATURE SURVEY

In the field of Natural Language processing, there are lot of approaches have been proposed for concept extraction BorisGelfand, MarilynWulfekuhler ,WilliamF.PunchIII [1] describe a system for extracting concepts from unstructured text without using NLP. It is achieved by identifying relationships between words in the text based on a lexical database and identify the conceptual groups. The word relationships are used to create directed graph, called a SemanticRelationshipGraph (SRG) which represent the relationships between word senses,used to identify the individual concepts.

Philipp Cimiano, Andreas Hotho, Andreas Hotho[12] present the automatic acquisition of taxonomies or concept hierarchies from a text corpus. The approach is based on Formal Concept Analysis (FCA), a method used for the analysis of data, i.e. for investigating and processing the information. FCA produces a lattice that is converted into a concept hierarchy.

Yuefeng Liu ,Minyong Shi (2016)[16] Propose a method to extract ontology concepts from multiple text of the same type. This method uses mutual information and document frequency. N-gram algorithm is used to generate a set of candidate phrases. The statistics and the rules to screen for the concept of ontology from candidate phrases.

Paul M. Ramirez and Chris A. Mattmann [11] present the Automatic Concept Extraction (ACE) algorithm, which can aid users performing searches using search engines.

Qing Yang ,Kai-min Cai , Yan LI ,Rui-qing Liu(2010)[13] proposed an improved area concept extraction algorithm .Association rule algorithm is used to obtain the similarity between the sememes, which is used to find the similarity between area concepts.

Jon Patrick, Min Li(2010)[5]describe the cascaded approach with two machine learners such as CRF(Conditional Random Fields) and SVM(Support Vector Machine).CRF identify the Named Entities and SVM classify the relationship between those entities. Rule-based engine identify, whether or not the medication was under a medication heading in the report.

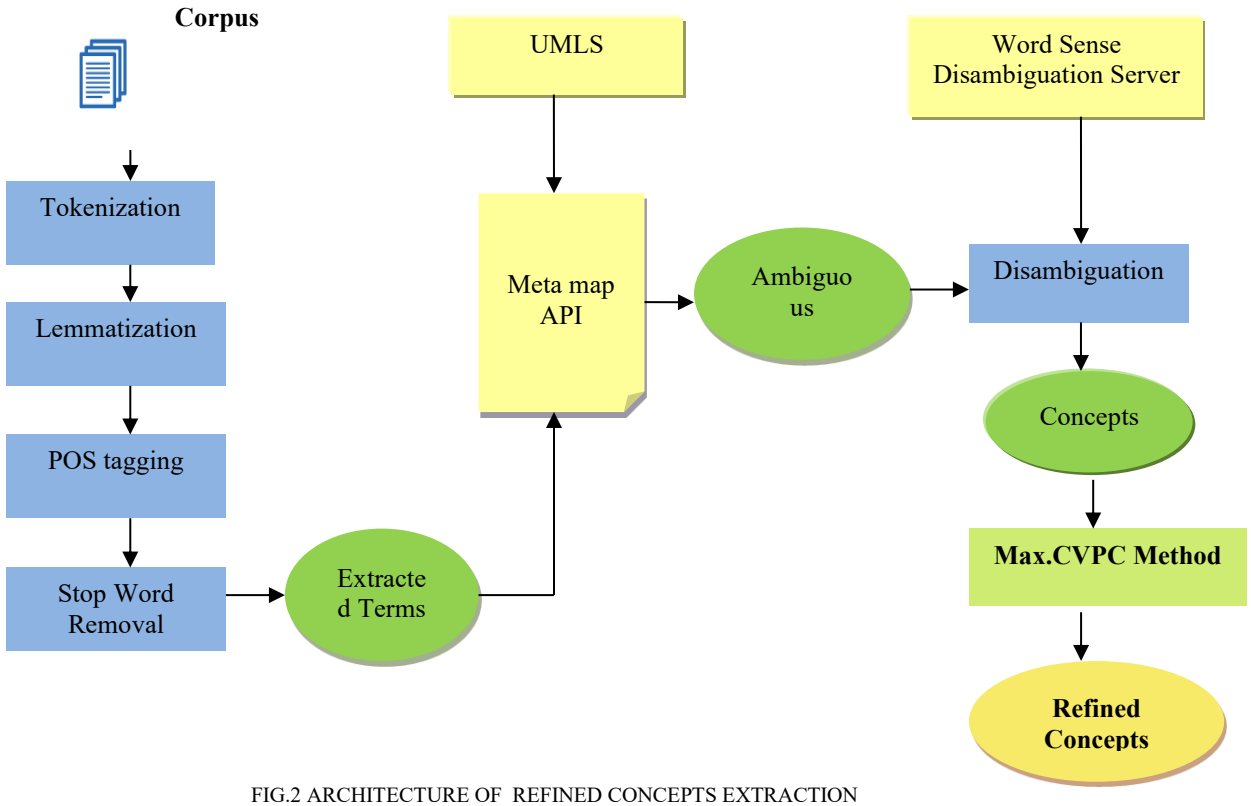

FIG.2 ARCHITECTURE OF REFINED CONCEPTS EXTRACTION

Raghavendra Chalapathy ,Ehsan Zare Borzeshi,Massimo Piccardi(2016)[14]proposed bidirectional LSTM(Long Short-Term

Memory) with CRF(Conditional Random Fields) decoding to employ contemporary neural network with the off-the-shelf word embeddings.

Luca Soldaini,Nazli Goharian(2016)[8] present QuickUMLS:a fast, unsupervised, approximate dictionary matchingalgorithm for medical concept extraction.

III. METHODOLOGY

Our proposed approach improves the quality of concept detection in simple and easiest way. Natural Language processing techniques are used to analyze, interpret and manipulate human language for the purpose of achieving human-like language processing. The input of NLP method is taken from MEDLINE/ PubMed resources. In this approach, NLP is used to perform tokenization, Lemmatization, part-of-speech tagging (POS) and stop words removal. These techniques are used for preprocessing the text and to get the proper biomedical terms.

Tokenization is the process that splits the artifacts into tokens .Stemming [9]is the process of converting or removing inflected form to a common word form(i.e.) it chop off the end or beginning of the word. In some cases, stemming does not produce the appropriate root word. For example,' studied' and 'studies' are changed to root term as 'studi'.

Form	Suffix	Stem
metabolites	es	metabolit
mediated	ed	mediat
strategies	es	strategi

TABLE 1:STEMMING

To overcome the problem of retrieving inappropriate biomedical terms, we go for lemmatization technique instead of stemming ,to get the accurate biomedical terms.

Lemmatization algorithm uses lexical knowledge bases to remove the inflectional endings and return the root or dictionary form of a word, which is known as the lemma. It depends on correctly identifying the intended part of speech and meaning of a word in a sentence.

Part-of-speech tagging assigns unambiguous lexical categories to each word such as nouns, verbs, adjectives, and others. POS Tagging looks for relationships within the sentence and assigns a corresponding tag to the word.

Form	**Suffix**	**Lemma**
metabolites	es	metabolite
mediated	ed	mediated
strategies	es	strategy

TABLE 2:LEMMATIZATION

```
/G/public_mm_win32_main_2014/public_mm/bin/SKRrun.14
/G/public_mm_win32_main_2014/public_mm/bin/metamap14.BINARY.x86-win32-nt-4 --lexicon db -Z 2014AA --
sldi -y
Berkeley DB databases (USAbase 2014AA strict model) are open.
Static variants will come from table varsan in
g:/public_mm_win32_main_2014/public_mm/DB/DB.USAbase.2014AA.strict.
Derivational Variants: Adj/noun ONLY.
Variant generation mode: static.
Established connection $stream(54874536) to TAGGER Server on localhost.
Established connection $stream(54874680) to WSD Server on localhost.

metamap14.binary.x86-win32-nt-4 (2014)

Control options:
  composite_phrases=4
  lexicon=db
  mm_data_year=2014AA
  sldi
  word_sense_disambiguation
Processing 00000000.tx.1: :95369245 IL-2 gene expression NF-kappa B activation CD28 require reactive oxygen
production 5-lipoxygenase

Phrase: :

Phrase: 95369245 IL-2 gene expression NF-kappa B activation CD28
Meta Mapping (618):
   654   IL-2 Gene (IL2 wt Allele) [Gene or Genome]
   623   Expression (Expression procedure) [Therapeutic or Preventive Procedure]
   654   NF-kappa B [Amino Acid, Peptide, or Protein,Immunologic Factor]
   623   Activation (Activation action) [Activity]

Phrase: require

Phrase: reactive oxygen production 5-lipoxygenase
Meta Mapping (766):
   637   Reactive [Qualitative Concept]
   637   Oxygen [Element, Ion, or Isotope,Pharmacologic Substance]
   637   production [Occupational Activity]
   804   Lipoxygenase [Amino Acid, Peptide, or Protein,Enzyme]
```

FIG.3 EXAMPLE OF UMLS CONCEPTS

The most common words or unwanted words in the biomedical corpus such as “the”, “a”,”an”, “for”,” be” etc are removed by stop word removal processing.

Text with stop words	Text without stop words
therapeutic strategy and the development	therapeutic strategy development
ROI formation which then induce IL-2	ROI formation induce IL-2

TABLE 3: STOP WORD REMOVAL

After removing the noise of the corpus, terms extracted are further processed to acquire the concepts. The biomedical concepts are retrieved with the help of UMLS.

Unified Medical Language System (UMLS) is a broad list of biomedical terms, which is used for developing applications in biomedicine and healthcare.

UMLS Knowledge Sources such as Metathesaurus contains huge mass of biomedical concepts from various source vocabularies. MetaMap maps the terms which are extracted

using NLP technique to identify the biomedical concepts. As part of this mapping process, MetaMap tokenizes text into sections, sentences, phrases, terms, and words.

It maps the noun phrases of the text to the best matching UMLS concept.

Concept ID and Concept Name are retrieved with ambiguous Concepts (i.e.) Metamap maps two or more concepts to recognized terms in the document. Disambiguation process is done by using Word Sense Disambiguation (WSD) server that determines which concept is the best choice for the entity using context in which the entity occur[15].

Although there is lot of duplication and unnecessary concepts such as functional concept, spatial concept, temporal concept etc in the retrieved results .To refine those concepts we proposed Max.CVPC method.

IV. PROPOSED WORK

Our proposed work is Hybridization of Maximum C-Value of Permutationof UML Concepts approach, where we extract the concepts in simplest and easiest way and also improve the accuracy of concepts.

After Preprocessing and Metamapping, normally successive sequence of words are extractedusing N-gram technique.It is mainly used to analyze the frequency of concepts. N-gram[3] is a sequence of N words where Unigrams are single word; Bigram is a sequence of two words, Trigram is a sequence of three words[12].

UML Concepts	Unigram	Bigram	Trigram
protein tyrosine kinase activity	'protein', 'tyrosine', 'kinase', 'activity'	'protein tyrosine', 'tyrosine kinase', 'kinase activity'	'protein tyrosine kinase', 'tyrosine kinase activity'
il 2 inducible enhancer	‘il’,’2’, ’inducible’,’enhancer’	‘il 2’, ’2 inducible’, ‘inducible enhancer’	‘il 2 inducible’, ’2 inducible enhancer’

TABLE 4:UMLS CONCEPTS BASED ON N-GRAM

C-Value is computed for these N-grams splitted concepts to find the dominant concepts.The C-value method combines linguistic and statistical information [6]. Linguistic information is the use of a regular expression as linguistic patterns, and the statistical information is the value assigned with the C-value measure based on the frequency of terms to compute the termhood (i.e.) the association strength of a termis used to extract domain concepts. Here the C-value method aims to improve the extraction of long terms as well as for extracting multi-word terms.

$$\text{C-value}(A) = \begin{cases} w(A)*f(A) \text{ if } A \nsubseteq \text{nested} \\ w(A)*\left[f(A) - \frac{1}{|S_A|} * \sum_{b \in S_A} f(b) \right] \\ \text{Otherwise} \end{cases}$$

Otherwise

Where A is the candidate term, w(A) = log2(|A|),|A| the number of words in A, f(A) the frequency of A in the document, S_A the set of longer terms that contain A and $|S_A|$ is the number of terms in S_A, f(b) the frequency of longer term that contain A in the document. C-value uses either the frequency of the term if the term is not included in other terms, or decreases this frequency if the term appears in other terms, based on the frequency of those other terms, Here the weightage increases based on the sub terms occurred in the multi word terms.

The N-gram Concepts are filtered with Maximum of C-Value from among the UML concepts (i.e)Dominant Concepts are extracted based on more weightage. The Precision is increased by filtering those dominant concepts with the threshold value.

We proposed the Max.CVPC method to refine the concepts with more weightage by finding the probability of concept using permutation.

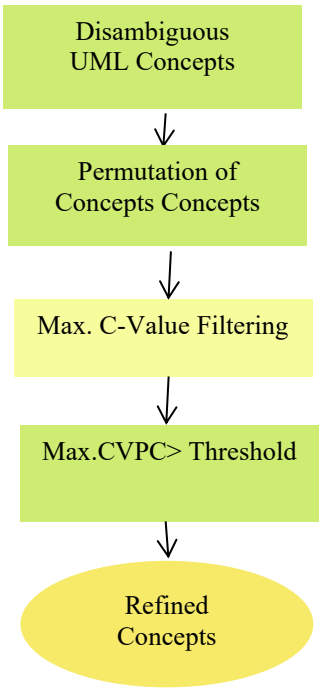

FIG.4 MAX. CVPC APPROACH WORKFLOW

Instead of N-gram technique we use the Permutation of word sequence to find the maximum possibility of concepts and then filtered by finding the maximum C-Value for those concepts.

UML Concepts	Permutation of UML Concepts
nf kappa b	[('nf',), ('kappa',), ('b',)] [('nf', 'kappa'), ('nf', 'b'), ('kappa', 'nf'), ('kappa', 'b'), ('b', 'nf'), ('b', 'kappa')] [('nf', 'kappa', 'b'), ('nf', 'b', 'kappa'), ('kappa', 'nf', 'b'), ('kappa', 'b', 'nf'), ('b', 'nf', 'kappa'), ('b', 'kappa', 'nf')]
epithelial cell	[('epithelial'), ('cell')] [('epithelial', 'cell'), ('cell', 'epithelial')]

TABLE 5:UMLS CONCEPTS BASED ON PERMUTATION

The set of concepts are again refined with the threshold value which increase the precision value compared with N-gram concepts.

Algorithm: Extracting Refined Concept

Input: Biomedical Corpus

Output: Refined Biomedical concepts

1. Preprocess the corpus with POS, Lemmatization and stop word removal.
2. For each sentence in sentences
    a. Terms=invoke Metamap in Sentence
3. End
4. For each entity in Terms
    a. Concepts=entity after performing disambiguation
5. End
6. For each concept in concepts
    a. Generate Permutation of concept
    b. Concept refinement by taking Maximum C-value (Permutation of Concept)
    c. If Max.CVPC>threshold

        Refined Concept=concept
7. End

V. EXPERIMENTAL RESULTS

The experiments with Max.CVPC approach were conducted to confirm its feasibility. This approach is implemented with Python. The evaluation was performed on 2,000 MEDLINE abstracts.

We compare the Max.CVPC method with Max-Value of Bigram and Trigram concepts. The Unigram concepts are not considered because when applying C-Value, the concepts obtain no weightage because there are no sub terms.

We selected the GENIA corpus[4]which contains 2000 abstracts totaling 4,20,000 words selected from 'National Library of Medicines' MEDLINE database for the biological domain. This was created by extracting only the main textual content from the HTML pages and ignoring formatting or navigational elements.GENIA Corpus is annotated with a subset of the substances and the biological locations involved in reactions of proteins, based on a data model of the biological domain, in XML format, which is used as Ground truth.

The Original concepts obtain from Ground truth i.e from GeniaCorpus.XML is 33298.After Metamapping, we obtain 14274 predicted concepts. The Max.C-Value of Bigram filter 6279 concepts and Max.C-Value of Trigram filter 2066 concepts. While the Max.CVPC extract 10223 dominant concepts.

To evaluate the proposed concept detection, we use precision, recall, and F-measure as evaluation parameters.

	Precision	Recall	F-Measure
Bigram+Max C-value	70.74%	65.68%	72.93%
Trigram+Max C-value	74.09%	66.73%	75.15%
Permutations+Max C-value	93.61%	78.11%	76.85%

TABLE 5:EVALUATION OF CONCEPT EXTRACTION

Precision is the ratio of the number of true positive concepts annotated over the total number of predicted concepts annotated as in Eq. (1).It denotes the number of correct concepts from the predicted result .

$$\text{Concept precision} = \frac{\text{Correct concepts}}{\text{Total concepts in the predicted result}} \quad (1)$$

The result is graphically depicted as:

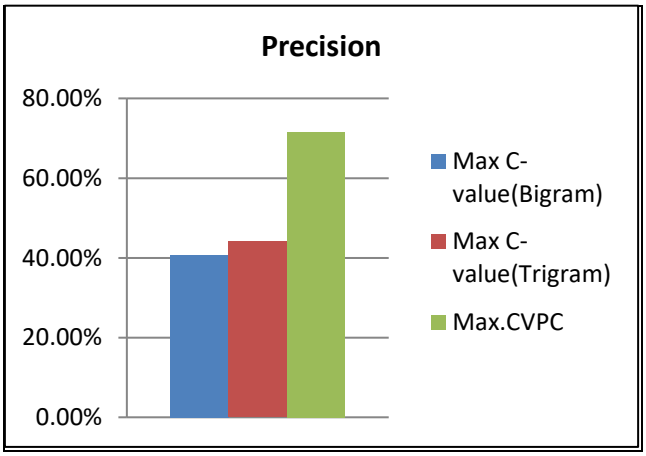

Fig.5 Comparison of Precision between Max.C-value(N-gram) and Max.CVPC approach

Recall is the ratio of the number of true positive concepts annotated over the total number of true positive concepts set as in Eq. (2). It denotes the number of correct concepts from the original result.

$$\text{Concept Recall} = \frac{\text{Correct concepts}}{\text{Total concepts in the original content}} \quad (2)$$

The result is graphically depicted as:

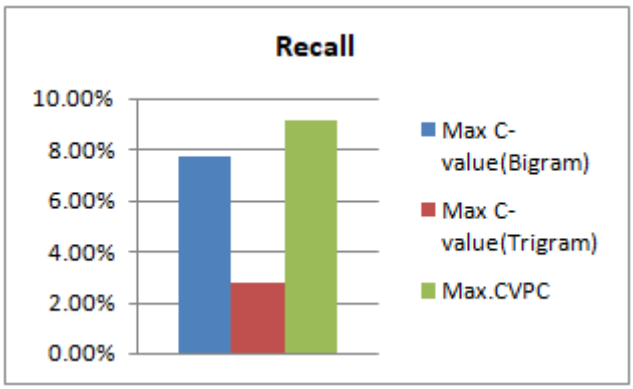

Fig.6 Comparison of Recall between Max.C-value(N-gram) and Max.CVPC approach

F-measure is the harmonic mean of precision and recall as in Eq. (3).

$$\text{F-Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

The result is graphically depicted as:

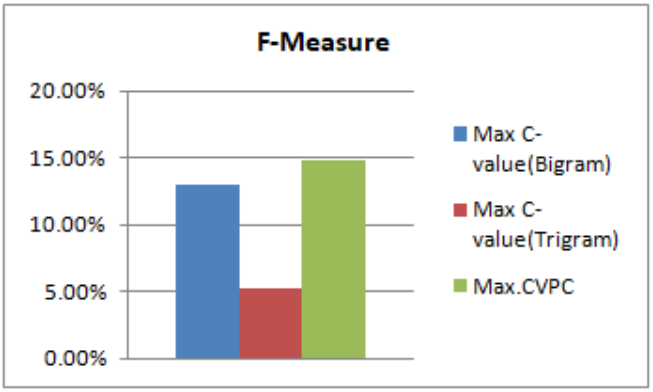

Fig.7 Comparison of F-Measure between Max.C-value(N-gram) and Max.CVPC approach

VI. Conclusion

Ontology is the cornerstone of the semantic web vision. It provides a common and shared understanding about concepts in a specific domain, reuse domain knowledge, and data interoperability. However, the manual ontology construction is a complex task and is very time consuming. Therefore, we presented a fully automated concept extraction method which minimize and downsize requirements and complexity, and improve the accuracy of ontology generation. Ontology is not used only for better search, data interoperability, and presentation of content, but more importantly it represents the foundation of future innovative ways to manage dormant content assets and transform the Web of document to Web of Data.To build quality semantic metadata we need dominant or refined concepts to improve the accuracy. Our proposed work extract those dominant concept and also improves the precision.

VII. Future Work

Our future work includes an extension of this work to support non-biomedical domain ontology generation. In addition, we concentrate on semantic enrichment in

automatic ontology learning and to further improve F-measure of concepts.

REFERENCES

[1]BorisGelfand, MarilynWulfekuhler ,WilliamF.PunchIII,"AutomatedConceptExtractionFromPlainText",1998 .

[2]Buitelaar,P.,Cimiano,P andMagnini,B, "Ontology learning from text: an overview. In: Ontology Learning from Text: Methods, Evaluation and Applications,Amsterdam,IOSPress, 123, 3–12,2005.

[3]Daniel Jurafsky & James H. Martin," Speech and Language Processing",chapter 4, Draft of September 1, 2014.

[4]Jin-Dong Kim , T. Ohta, Y. Tateisi and J. Tsujii , "GENIA corpus—a semantically annotated corpus for bio-textmining", BioInformatics Vol. 19 Suppl. 1 2003, pages i180–i182, DOI: 10.1093/bioinformatics/btg1023.

[5]Jon Patrick, Min Li,", Journal of the American Medical Informatics Association" ·,September 2010.

[6]K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: the c-value/nc-value method. International Journal on Digital Libraries, 3(2):115–130, 2000.

[7]K.Karthikeyan,V. Karthikeyani,"Ontology Based Concept Hierarchy Extraction of Web Data", Indian Journal of Science and Technology, Vol 8(6), 536–547, March 2015.

[8]Luca Soldaini,Nazli Goharian ,"QuickUMLS: a fast, unsupervised approach for medical concept extraction ",MedIR Workshop at SIGIR 2016.

[9]Lovins JB. Development of a stemming algorithm; 1968. p. 22–31.

[10]Muhammad Nabeel Asim, Muhammad Wasim, Muhammad Usman Ghani Khan, Waqar Mahmood and Hafiza Mahnoor Abbasi," A survey of ontology learning techniques and applications",2018.

[11] Paul M. Ramirez and Chris A. Mattmann," ACE: Improving Search Engines via Automatic Concept Extraction".

[12] Philipp Cimiano, Andreas Hotho, Andreas Hotho," Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis", Journal of Artificial Intelligence Research 24 (2005) 305–339 ,Aug 2005.

[13] Qing Yang ,Kai-min Cai , Yan LI ,Rui-qing Liu," An Area Concept Extraction Algorithm Based on Association Rule", International Conference On Computer Design And Appliations (ICCDA 2010),2010.

[14]Raghavendra Chalapathy ,Ehsan Zare Borzeshi,Massimo Piccardi," Bidirectional LSTM-CRF for Clinical Concept Extraction", Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP),December 2016.

[15] Wanda Pratt, Ph.D, Meliha Yetisgen-Yildiz, M.S," A Study of Biomedical Concept Identification: MetaMap vs. People", AMIA 2003 Symposium Proceedings − Page 529.

[16]Yuefeng Liu ,Minyong Shi, "Domain Ontology Concept Extraction Method Based on Text",IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS),2016.

Authors Profile

R.Manimala is currently pursuingher part-time research in the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli. She completed her M.Phil degree in the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli and Master Degree in Computer Applications from SCAD college of Engineering and Technology, Cheranmahadevi. She is working as an Assistant Professor at Sri Paramakalyani College, Alwarkurichi, Tamilnadu. Her research interest includes Data mining and big data.

Dr.G.Muthulakshmi is an Assistant Professor in the Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli. She received her B.E degree in Computer Science and Engineering at PSR Engineering College, Viruthunagar and M.E degree in Computer Science and Engineering from Manonmaniam Sundaranar University, Tirunelveli. She received her doctorate in Computer Science and Engineering from Manonmaniam Sundaranar University, Tirunelveli. She has 11 years of teaching experience and 9 years of research experience. Her areas of interest are Digital Image Processing, Data Mining, and Neural Networks.