ISSN 2277-3061

Semi-Automated Ontology building through Natural Language Processing

Jaytrilok Choudhary, Deepak Singh Tomar
Department of CSE, MANIT Bhopal (M.P.), India, 462003
Department of CSE, MANIT Bhopal (M.P.), India, 462003

ABSTRACT

Ontology is a backbone of semantic web which is used for domain knowledge representation. Ontology provides the platform for effective extraction of information. Usually, ontology is developed manually, but the manual ontology construction requires lots of efforts by domain experts. It is also time consuming and costly. Thus, an approach to build ontology in semi-automated manner has been proposed. The proposed approach extracts concept automatically from open directory Dmoz. The Stanford Parser is explored to parse natural language syntax and extract the parts of speech which are used to form the relationship among the concepts. The experimental result shows a fair degree of accuracy which may be improved in future with more sophisticated approach.

Indexing terms/Keywords

Ontology; Parsing; Knowledge representation; Ontology building.

Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATION JOURNAL OF COMPUTERS AND TECHNOLOGY
Vol. 13, No. 8
editorijctonline@gmail.com
www.ijctonline.com, www.cirworld.com

INTRODUCTION

Ontology is a very important discipline in the areas of natural language processing. The main use of ontology is for knowledge representation, organisation and its acquisition. It is backbone of semantic web. There are various ways to define "what ontology is?" From Artificial Intelligence point of view, ontology is defined as "*explicit specification of conceptualization*". Conceptualization is the abstract representation of a real world entity with the help of domain relevant concepts [1]. From knowledge-based systems point of view, it is defined as "*a theory of concepts/vocabulary used as building blocks of an information processing system*" [2]. A compositional definition of ontology is "*Ontology is a hierarchical organization of concepts along with relationship between them*". An ontology O can be represented as O := (C, R, D, H) where

- C : A set of concepts
- R : A set of relationships between concepts
- D: Domain name for which ontology has been constructed
- H : A hierarchy H = (V, E) defined as a simple directed graph where V is a set of nodes V that represents concepts of domain and E is a set of ordered pairs called edges $(C_1, C_2) \in E \subseteq \{ V \times V \}$. The direction of an edge (C_1, C_2) is from concept C_1 to concept C_2 specify that C_1 relates to C_2 by relationship r where $r \in R$.

In general ontology can be broadly classified into two categories: Natural Language Ontology and Domain Ontology [3]. Natural Language Ontology has language concepts and lexical relations between them. In general, these concepts are very large in size and do not require frequent updates. Word Net is an example of natural language ontology. Domain ontology represents knowledge of one particular domain only. For example, ontology of computer domain represents concepts of computer domain and relationship between these concepts only. This ontology presents detailed explanation of the domain concepts only. Therefore, it is also referred as 'vertical' ontology.

Ontology should be formal to become machine understandable and enable to share knowledge across the communities [1]. There are some ontology management tools available which provides the facilities and environments to build a new ontology [10]. These type of ontology construction is referred as manually ontology construction. The manual ontology construction is very easy. It requires lots of human efforts. So, It is very time consuming and costly. For correct ontology construction adequate domain knowledge is required. The efficient ontology can be constructed by domain experts only.

Now-a-days maximum research is on semiautomatic and automatic ontology construction to overcome the problems of manual ontology construction. Semi-automated ontology construction begins with small core ontology constructed by domain experts and learns the new concepts and relationships between concepts automatically using expert algorithms. in automated ontology construction, new concepts and relationship between these concepts are learnt automatically using expert algorithms. There are various approaches have been proposed to build automated and semi-automated ontology [4-8]. However, more work is required to construct ontology automatically with good accuracy.

In this paper, a method to build ontology in semi-automated way has been proposed. The basic idea motivating is to use the information available in Web to develop a domain ontology in semiautomatic manner. The two main aspects are covered in this paper first automatic extraction of the domain concepts from web and second to find relationship between these concepts automatically. The domain concepts are extracted from open directory project Dmoz and natural language processing technique is used to find the relationship between concepts. The rest of the paper is structured as follows. Section 2 covers the related work done for ontology building. The proposed ontology building framework to develop the domain ontology is presented in section 3. Section 4 covers the ontology building algorithm. Section 5 presents the experimental results and finally, concluded our work.

RELATED WORK

There are various methods for ontology building has been proposed, few important are discussed here.

Mei-ying Jia et. al.[4] has proposed an automated ontology building method. This method is not completely automated. It uses open thesaurus and Military Intelligence database. The thesaurus gives various classes information for the ontology and the database gives the instances. There are three types of relationships are used between concepts to form ontology: is_a, part of (has) and synonymy. At the end, Protégé open source editing tool is used to represent ontology that provides a friendly interface for users.

H. Kong et. al. [5] gave the algorithm to build the ontology automatically. The frame ontology is constructed from Word Net concepts and existing knowledge data. The ontology construction algorithm works into two parts. In first part, it is to make the possibility for building the automatically ontology using the frame ontology constructed from Word Net concepts that are the standard structured knowledge data. In second part, domain experts uses specific input data to make the ontology more complete. This algorithm is not completely automatic, here relationship are limited to Word Net relationship only and initial core concepts are taken from WorldNet.

J. Wang et. al. [6] proposed an ontology learning algorithm that learns ontology instances using rule-based information extraction. The factors of the instances are automatically extracted using the definition of domain ontology. The rule generation technique is used for Information Extraction. The rule generation algorithm applies supervised learning with bottom-up strategy and uses a heuristic method to decide rule generalization path. Laplacian* formula is used to evaluate the performance of rules.

$$Laplacian^{*} = (e_s + e_t + 1)/(n + 1) \qquad (1)$$

Where n is number of extractions made on the training set, e_s is the number of substitutions and e_i is the number of insertions.

Q. Yang et. al. [7] gave an Ontology building algorithm that combines stable domain concept extraction method and personalized recommendation with concept extraction. It uses machine learning for extraction of domain concept. Domain concept are extracted using recommendation study. This technique improves the accuracy of the concept extraction and the stability but there are still many issues to be handling like relationship learning.

Wu yuhuang et. al. [8] gave a web based ontology building model. This technique concerns realizing an automatic extraction of ontology from the Web page and discovering the pattern and the relationships of the ontology concepts from the Web page data. It extracts Web ontology semi-automatically through the analysis of Web page collection in the identical application domain.

Wen Zhou et. al. [9] proposed a semi-automatic ontology building technique that starts from small core ontology built by domain experts and learns the new concepts and relationships from Word Net and event-based natural language processing technologies to construct the target ontology. Relationship learning for ontology is based on event extraction that finds out the verb relationship between concepts. This method is completely based on Word net to discover relationship between concepts and verb extracted during natural language processing.

ONTOLOGY BUILDING FRAMEWORK

The overall architecture of ontology building framework is shown in Fig. 1. The large corpus of various domains like Art, science, computer and sports is collected from the Dmoz Open Directory. The preprocessing is required to extract domain specific concepts from large corpus. Stemming is used to obtain the root word. The extracted domain oriented concepts are stored in the database. An algorithm is applied to set the relationship among the concepts. The ontology building steps are described below.

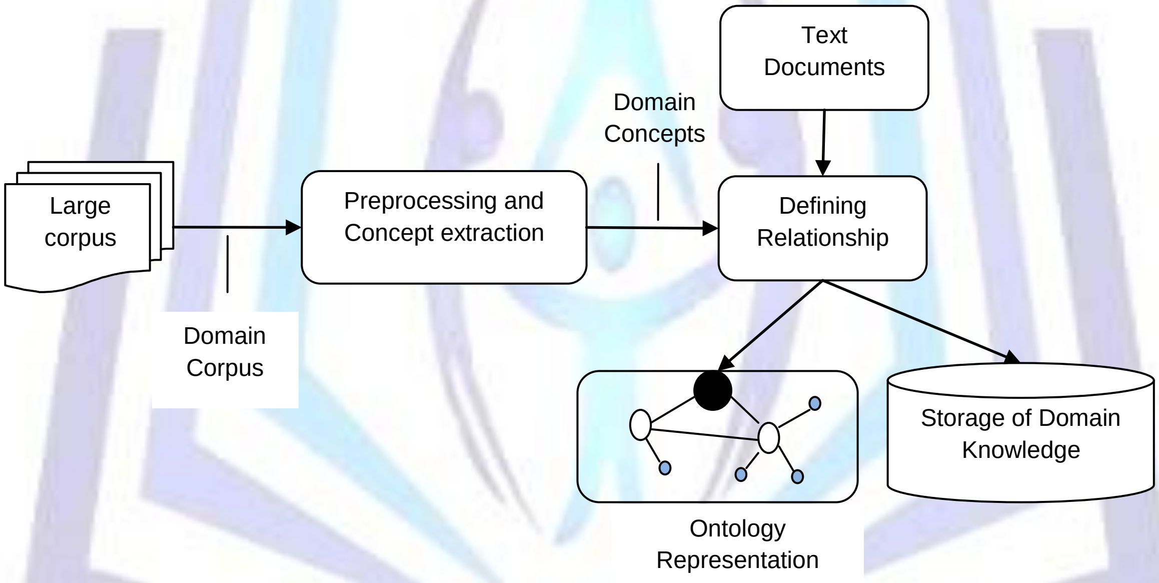

Fig 1: Architecture of Ontology Building Framework

Document Collection from Large Corpus

The large corpus is collected from Dmoz Open Directory Project that is the most comprehensive human-edited directory in the WWW. It contains information about various domains like art, business, computer, sports and so on. It is constructed and maintained by a vast and global community of volunteer editors. The corpus present on Dmoz is in RDF (Resource Description Framework) format. For example, all the domains are attached under top label then corpus of every domain in RDF file as shown below

```
<Topic r:id="Top">
   <catid>2</catid>
   <d:Title>Top</d:Title>
   <lastUpdate>2010-02-16 08:43:34</lastUpdate>
   <d:Description></d:Description>
   <narrow r:resource="Top/News"></narrow>
   <narrow r:resource="Top/Science"></narrow>
```

```
    <narrow r:resource="Top/Business"></narrow>
    <narrow r:resource="Top/Health"></narrow>
    <narrow r:resource="Top/Computers"></narrow>
    <narrow r:resource="Top/Sports"></narrow>
    <narrow r:resource="Top/Arts"></narrow>
………………………………………………………………………………………………
</Topic>
```

Only domain oriented RDF structure is extracted from large corpus. If the working domain is Computer, all the Computer oriented RDF structures are extracted.

```
<Target><related >
 <"Top/Computers/Mobile_Computing/Wireless/Software"/>
<"Top/Computers/Hardware/Mouse/">
<"Top/Computers/Internet/Protocols/">
<"Top/Computers/Data_Communications/Internet">
<"Top/Computers/Data_Communications/Wireless">
<"Top/Computers/Software/System software"/>
<"Top/Computers/Software/Application software"/>
<"Top/Computers/Software/Operating System"/>
<"Top/Computer/Software/Automation/Manufacturing ">
<"Top/Computers/Software/ERP">
<"Top/Computers/Software/Graphics/Color_Management">
<"Top/Computers/ Firmware/Graphics/Fonts"/>
………………………………………………………………………………………………
</related></Target>
```

Preprocessing and concept extraction

The preprocessing and concept extraction is divided into two parts: RDF (Resource Description Framework) parsing and stemming.

RDF Parsing: The extracted corpus is present in RDF file format. So, first it is needed to parse it to collect domain oriented concepts. In parsing, each line is scanned to find domain word first. Now, two level hierarchy words have been selected after from domain word. For computer domain, concepts are shown below in Fig 2.

All these words belong to computer domain concepts. For example: *Software, hardware, Data communication, mobile computing, application software, system Software* and so on. After this step, all the domain oriented topics, sub topics and concepts are extracted from above RDF structure.

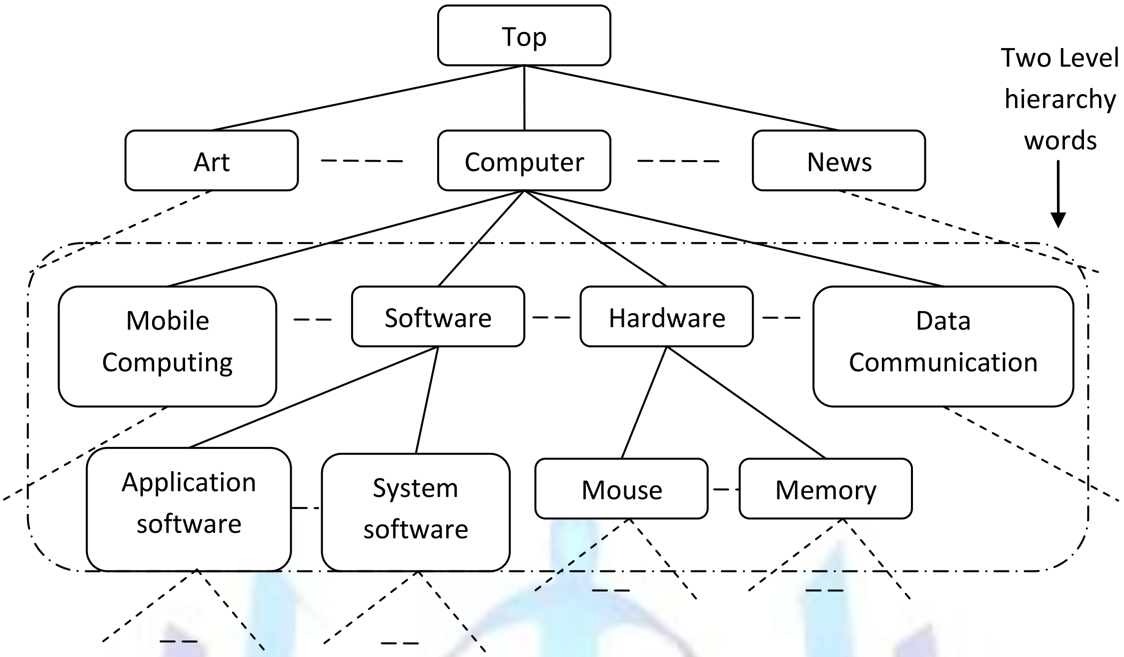

Fig 2: Domain oriented concept extraction

Stemming: A group of words where words in the group are small syntactic variants of one another may share the same word stem. So, it is useful for the ontology learning system to identify such group of words and collect only the root word stem per group. For example, the groups of words: *computation, computing, computes* shares a common word stem, *compute*, and must be viewed as the same word for different occurrences.[11]

Defining Relationship among the concepts

Defining Relationship is a challenging part of ontology building process. Most of the ontology keeps a few relationships between the concepts like 'Is-a' and 'Part-of' [9]. However, the relationship list should be broad enough to cover most important forms so that the ontology can be utilized widely by different applications.

The Stanford Parser is used for defining relationship among the concepts. It is a natural language parser that works out the grammatical structure of sentences to find out which groups of words go together (as phrases) and which words are the subjects, object or a verb. It takes the English text as input and outputs the corresponding POS (parts of speech) tags for each word. It uses lexical and syntactic annotations to denote the part of speech of the terms; for example, NN denotes a proper noun, VB denotes a verb, NP denotes a noun phrase etc as shown in Table 1.

Table1. English sentence and its POS

Tasks	**Examples**
Sentence	Computer operated by software
POS	('Computer','NN')('operated','VBD')('by','IN')('software','NN')

To extract relationship between concepts, collected text documents are processed. Sentences are parsed and semantic of sentences are learned using Stanford Parser. Different rules are formed to assign relationships between concepts.

The relationship learning algorithm works as follows:

Step 1: Process the document and extract sentences one at a time; ignore sentences that do not contain two nouns.

Step 2: Check, at least one of the noun should be a concept. If one noun (concept) is already present in the ontology then add second noun (concept) in the ontology and if both the nouns are present then try to find out the relationship between these two noun words (concepts) from that sentence.

Step 3: Let C1 and C2 are two concepts in step 2. Preserve their order and all the words in between them.

Step 4: Infer the relationship (R) with the help of Verb, Preposition and Adjective word that occur along with these concepts.

Step 5: Add relationship between these two concepts (C1, C2, R) in the ontology.

Storing and representation of ontology

After defining the relationship between the concepts, concepts along with relationships are stored in the database as given in Table 2.

Table2. Concepts and relationship between concepts

Concept1	Concept2	Relationship
Computer	Software	Is_Operated_by
Computer	Hardware	Has
Operating system	Software	Is_a
Operating system	Hardware	Operates
....	....	....

Finally, Ontology is represented in graph form as shown in Fig. 3.

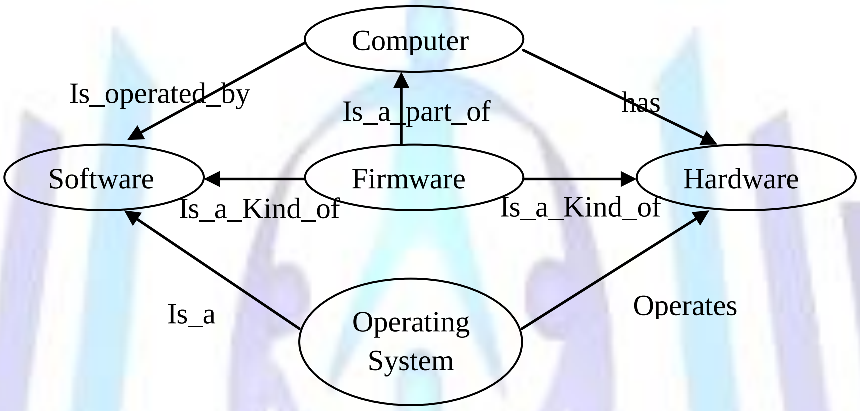

Fig 3: Graphical representation of ontology

ONTOLOGY BUILDING FRAMEWORK

Ontology learning algorithm is divided into three modules: sentence parsing and analysis, Verb checking for relationship and add concepts and relationship in database. Each module is described as follows:

Sentence parsing and analysis

Extract each sentence from the input text document. Parse each sentence using Stanford Parser to extract all the parts of speech. Now analyze parsed sentence to get relationship between concepts and new concepts.

```
Step 1: input a text document
Step 2: for each sentence S of document do
Step 3: PS : = Parse(S);          //sentence after parsing PS
Step 4: Count no. of nouns in PS.
Step 5:   If less than two then ignore sentence and Goto Step2.
Step 6:   If exactly two noun N1 and N2 then
              a.  If N1 and N2 are consecutive or connected using conjunction like AND, OR then ignore that sentence
                  Goto step 2.
              b. else
                  Verb_Checking(PS, N1, N2);
Step 7:   If more than two noun let n nouns then
              a. If all the nouns are consecutive or connected using conjunction like AND, OR then ignore that
                 sentence Goto step 2.
              b. if n-1 nouns are connected using conjunction like AND, OR (e.g. N1   AND N2 AND …AND Nn-2 OR
                 Nn-1 V Nn) then
```

```
                Verb_Checking(PS, N1, Nn);
                 …………….
                Verb_Checking(PS, Nn-1, Nn);
           c.  if some nouns are consecutive then
               Group them and treat that group as a single noun
               Goto step 7
```

Verb checking for relationship

During verb checking, the various cases to form verb as a relationship are handled.

Verb_checking(PS, N1, N2)

```
Step 1: count no. of verbs present in PS
Step 2: if exactly one verb V in it then
        a. If  V is 'was' or 'were' then
               Ignore that sentence return.
        b. else
           (i) if there is a preposition P or determiner D after V then add VP or VD as a relationship between N1 and N2
               add(N1, VP or VD, N2);   // e.g. Computer operated by software
           (ii) else
               add the verb V as relationship between N1 and N2
               add(N1, V, N2)        // e.g.  Computer has hardware.
Step 3: if exactly two verbs V1 and V2 then
            a. if V1 and V2 are consecutive then
                   ignore V1 and add the verb V2 as relationship between N1 and N2 add(N1, V2, N2)
                                                   // e.g.  Computer can run program.
            b. if V1 and V2 are connected using conjunction like  AND, OR then
                 add V1 and V2 both as a relationship between N1  and N2
                 add(N1, V1, N2); add(N1, V2, N2);
                                              // e.g. Software operates and configures hardware.
Step 4: If more than two verbs ignore that sentence S.
```

Add concepts and relationship in database

Add concept module stores relationship between concepts and concepts in database.

add(N1, V, N2)

```
Step 1: N1 and N2 are two concepts.
Step 2: Check one of the noun should be a concept.
Step 3: If one noun (N1) is already present in the ontology then add second noun (N2) in the ontology and relationship between N1 and N2 is V
Step 4: If both the nouns are present then
             add V as a relationship between N1 and N2
Step 5: else
         return;
```

EXPERIMENTAL RESULTS

The ontology has been developed for various domains like Art, science, computer and sports. The precision and recall metric are widely used in the field of Information extraction to evaluate the effectiveness of domain concept extraction [4].

$$\text{Precision} = A_{\text{Concept}} / T_{\text{Concept}} \quad (2)$$

$$\text{Recall} = A_{\text{Concept}} / D_{\text{Concept}} \quad (3)$$

Where A_{Concept} = Total number of concept extracted accurately, D_{Concept} = Total number of domain specific concept and T_{Concept} = Total number of concept.

The experimental results of proposed ontology building algorithm have been compared with ontology learning model [7].

The experimental results of ontology learning model [7] are shown in Table 3.

Table3. Precision and Recall of various domains by ontology learning model [7]

Domain	T_{Concept}	D_{Concept}	A_{Concept}	Precision	Recall
Art	2650	2170	1430	54%	66%
Science	3590	2985	1956	54%	66%
Computer	4350	3870	2820	65%	73%
Sports	3275	2690	1948	59%	72%

The experimental results of proposed ontology building method are shown in Table 4.

Table4. Precision and Recall of various domains by proposed ontology building method

Domain	T_{Concept}	D_{Concept}	A_{Concept}	Precision	Recall
Art	2650	2170	1880	71%	87%
Science	3590	2985	2556	71%	86%
Computer	4350	3870	3350	77%	87%
Sports	3275	2690	2380	73%	88%

The Fig 4 shows the comparison of accurate concepts extracted between proposed ontology building and ontology learning model [7].

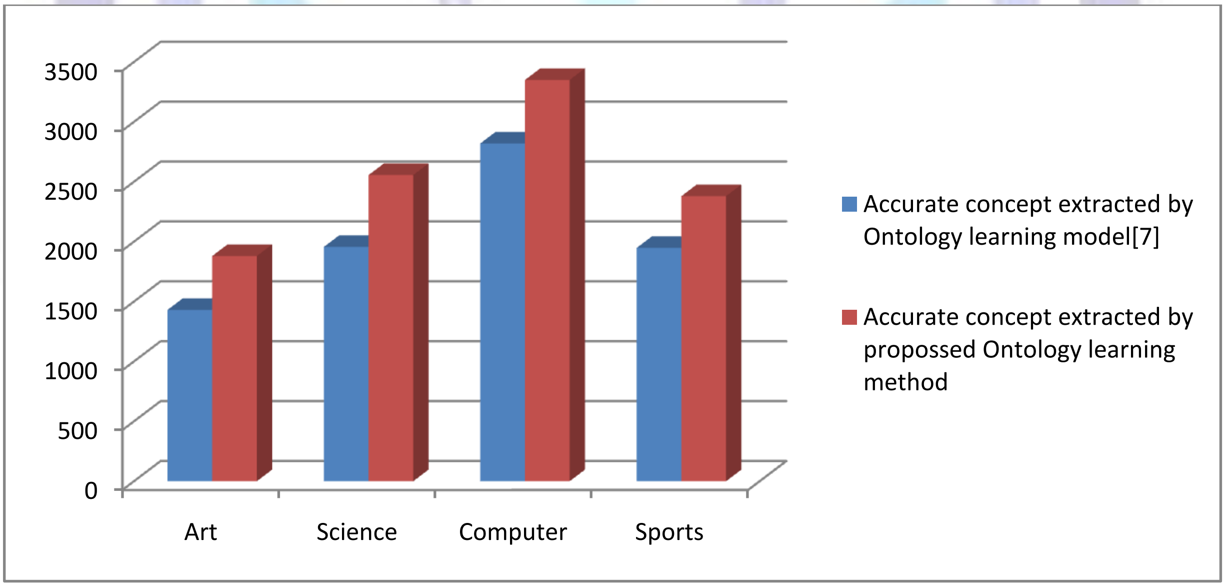

Fig. 4: comparison of accurate concepts extracted between proposed ontology building and ontology learning model [7]

The overall performance gain is calculated by equation (4).

$$Improvement\ \% = \frac{P-Q}{Q} \times 100\% \quad (4)$$

Where P : average number of accurate concepts extracted by proposed ontology learning method and Q : average number of accurate concepts extracted by ontology learning method [7].

The average concepts extracted by proposed ontology learning method and ontology learning method [7] are 2541.5 and 2038.5 respectively. Thus, the overall improvement achieved 23.52%.

CONCLUSION

In this paper, an ontology building methods has been proposed that builds ontology in semi-automated way. The domain concepts are extracted automatically from the large corpus Dmoz Open Directory. Each sentence of text documents is parsed through Stanford Parser and all the parts of speech are gathered to find relationship among the concepts. Finally, the ontology is stored in database and displayed in the graphical form so that users can understand the complete ontology at a glance easily and use it according to their necessities. The overall improvement of proposed method with respect to ontology learning method [7] is 23.52%.

REFERENCES

[1] Bhowmick, P. K., Roy, D., Sarkar, S. and Basu, A. 2010. A Framework for Manual Ontology Engineering for Management of Learning Material Repository. International Journal of Computer Science and Applications, 7, pp. 30 - 51.

[2] Gruber, T. R. 1993. A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition 5, 199–220.

[3] Jia, M., Yang, B., Zheng, D., Sun, W., Liu, L., and Yang, J. 2009. Automatic Ontology Construction Approaches and Its Application on Military Intelligence. In: Asia-Pacific Conference on Information Processing, pp. 348 – 351.

[4] Kong, H., Hwang, M. and Kim., P. 2006. Design of the automatic ontology building system about the specific domain knowledge. In: 8th International Conference on Advanced Communication Technology.

[5] Mizoguchi R., Vanwelkenhuysen, J. and Iked, M. 1995. Task ontology for reuse of problem solving knowledge. In: Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing.

[6] Omelayenko, B. 2001. Learning of Ontologies for the web: the analysis of existent Approaches. In: The international Workshop on web Dynamics held in conjunction with the 8th International Conference on Database theory.

[7] Porter, M. 1980. An algorithm for suffix stripping. Program, 14, pp. 130-137.

[8] Wang, J., Wang, C., Liu, J. and Wu., C. 2006. Information Extraction for learning of Ontology Instances. In: IEEE International Conference on Industrial Informatics.

[9] Yuhuang, W. and Yusheng, L. 2009. Design and Realization for Ontology Learning Model Based on Web. In: IEEE International Conference on Information Technology and Computer Science.

[10] Yang, Q., Cai, K., Sun, J. and Li, Y. 2010. Design Analysis and Implementation for Ontology Learning Model. In: 2nd International Conference on Computer Engineering and Technology.

[11] Zhou, W., Liu, Z., Zhao, Y., Xu, L., Chen, G., Qiang, W., Huang, M. and Qiang, Y. 2006. A Semi-automatic Ontology Learning Based on Word Net and Event-based Natural Language Processing. In: International Conference on Information and Automation.

Author’ biography with Photo

Jaytrilok Choudhary is currently Asst. Professor, in Maulana Azad National Institute of Technology (MANIT), Bhopal (India). He obtained his Bachelor of Engineering from Medicaps Institute of Technology and Management, Indore University Rajeev Gandhi Technical University Bhopal (India). He received his Masters degree in Computer Science and Engineering from Indian Institute of Technology, Roorkee (India). At present, he is pursuing Ph.D. from MANIT Bhopal (India).

Dr. Deepak Singh Tomar obtained his B.E., M.Tech and Ph.D. degrees in computer science and engineering. He is currently an assistant professor of computer science and engineering at NIT Bhopal, India. His research interests are in web mining and cyber security. He has published more than 35 papers and guided 23 M.Tech. Thesis.