ISSN:0975-9646

A Survey on Ontology Building Methodologies and Tools for Indian Languages

Pooja Jidge[1], Prof. Sharvari Govilkar[2]

Department of Computer Engineering, PIIT, India

***Abstract*—Ontology is one of the central areas in natural language processing (NLP) and artificial intelligence. Ontology is important for exploring the information for an efficient and useful management of knowledge and establishing communication between different entities such as human and application for understanding as well as reusing and sharing of knowledge. This paper reviews about the use of different ontology building approaches in NLP and their usage in Indian languages. The paper presents a survey of approaches for ontology building and its use in Indian regional language as well as challenges regarding Indian language usage of ontology mainly due to Indian languages being morphologically rich, resource constraint, lack of expertise and knowledge in building ontology.**

***Keywords*— Ontology engineering, domain, knowledge representation, information retrieval (IR), F-measure.**

I. INTRODUCTION

When surfing the web, the end users are increasingly in need of powerful tools that is capable of searching and interpreting the large amount of heterogeneous information present on internet. The data is found in various different formats. All these formats make it difficult for user to understand the relationship between the data. Ontology captures data in such a way that allows these relationships to be visible. Ontologies have the potential to significantly improve the process of information retrieval on web. The most acceptable definition of ontology was given by Tom Gruber. Gruber outline ontology as: "Ontology is a formal, explicit specification of a shared conceptualization" [1]

.

II. LITERATURE SURVEY

In this section we cite the relevant past literature for ontology building across various domain for English and different Indian languages. These techniques can be used for developing ontology to extract meaningful information from unstructured text.

Kong [2], implemented automatic ontology building system using WordNet. This system was helpful in reducing inconsistency and heterogeneity in ontologies. Kong mentioned the drawback of the existing ontology tools and suggested new methodology for automatic ontology building which consists of different steps. These steps when applied to ontology building tool it reduces semantic problem and developed the tree for the specific domain.

Saraswathi [3] proposed a system for information retrieval on festival domain. The system can be extended to any Indian language. The authors used ontological tree for inter- language conversion that allows user to query in their native language. Naïve algorithm was used for document search and page ranking algorithm in IR phase. It was noted from results that relevance of IR on document was improved for English by 40% and 60% for Tamil.

Panceras [4] present challenges and solution for building domain ontologies for indigenous languages. Several minority languages are morphologically rich, ambiguous and lack sufficient resources as well as knowledge and expertise in developing methods for building ontology. Also the ontology development cost is subjected to factors such as ontology maintenance, reuse, ontology quality and time required to build best ontology methods, analysis and controlling of ontology complexities is crucial.

K.R. Ananthapadmanaban [5] designed user profile ontology for Tamil Nadu tourism. By identifying the interest of user, the system suggest appropriate package of tourism for Tamil Nadu region.

S.M.Chaware [6] present approach for building ontology for grocery shop domain. Chaware presented pitfalls in the existing methodologies and proposed a system for ontology building. The author proposed ontology building approach algorithm that includes various different modules. The result obtained shows the whole, simple and easy creation of ontology from databases.

A.Kanaka Durga [7] presents a method for text categorization using ontology model for Telugu document. The ontology based classification included tokenization of document to obtain set of words then morphological analyzer is used to get root word. The retrieval model used is a vector space model. The disadvantage of classical bag of words model over efficient ontology model is presented.

Saraswathi [8] designed semi-automatic ontology tree which are created partially manual and completed dynamically. The semi automatic ontology tool improved the efficiency of IR relating to the user's query. Two parameters mainly recall and precision were used for testing the performance of a system

Brijesh Bhatt [9] proposed k partite graph learning algorithmic program for Indian languages that extract ontology from unstructured text. The algorithm showed improvement in precision without affecting the F-score. The proposed approach not only reduces the computation for ontology construction but also provides additional benefit of term filtering.

Sandeep [10] proposed an approach for semantic matching based on ontology for Hindi and Marathi language inference system. The ontology is built from

relational databases with translation/transliteration rules and query formation rules. The result obtained showed 20% and 18% precision for Hindi and Marathi users respectively. The recall obtained was 65% for Hindi user and 90% for Marathi users.

Iti [11] developed ontology for the health domain. This system can map similar concepts and relation available on different sources. The methodology included various phases such as specification, conceptualization, creation of instances and visualization. The ontology built included all the concepts and this can further be expanded for the future work.

James [12] proposed a graph model of domain ontology and used it for the text classification. The objective of document ontology graph is to carry text classification by matching the single document ontology with Domain ontology.

Rajveer Kaur[13] described the pre-processing phase for creation of ontology graph from Punjabi text documents. The preprocessing phase included various steps mainly removal of useless symbols, duplicate words and stop words. Punjabi dictionary and gazetteer list is used to identify the meaning of extracted term. These lexical resources can be used for developing NLP system in Punjabi language

III. ONTOLOGY BUILDING METHODOLOGIES

There are many ontology building methodologies suggested for various domains. With this methodology, one can effectively build the ontology with all possible user scenarios or simple and complex keyword. Some of the methodologies in building ontologies are as follow [14], [15], [16]:

A. Skeleton Methodology

A plan for a project along with activities can be represented as ontology. The steps are identifying the main purpose of the ontology, build the ontology and third, code it with proper language. It is simple but with limited scope.

B. Gruninger And Fox Methodology

Gruninger and Fox methodology is the result of the experience based on TOVE project. It is used to build a logical model of knowledge. This is not constructed directly, rather first it informal descriptions are made and based on them it is formalized.

C. Methontology Methodology

Methontology methodology will give the knowledge level ontology construction. The ontology development process is: Determine the tasks to be performed, determine the life cycle of ontology as number of stages, determine the techniques used in each activity.

D. Sensus Methodology

The SENSUS based method includes a series of terms are taken as seed. These seed terms are linked by hand to SENSUS where all the concepts in the path from the seed terms to the root of SENSUS are included.

E. WordNet Methodology

WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short, general definitions. It maintains the record of various semantic relations between these synonym sets.

IV. ONTOLOGY BUILDING PROCESS

A. *System Module*

The ontology building process includes various modules. The following is one of the approach or technique to build ontology for English as well as various Indian languages [14].

1) *User Interface:* The User interface is used to get a query from the user whose solution is to be found. The various sub modules under this are Language selection and User's query which includes selecting language of user choice and providing output in the same language.
2) *Parsing Module:* The parsing module will parse the complex keywords. Each keyword is taken as a input to the next module.
3) *Stemmer module:* Stemmer module stems the keyword by removing all stop words in order to find the root word. This root word will be given for translation.
4) *Translator Module:* This module will translate the entered stem keyword into English. Translation will be done on each parsed strings.
5) *Query Module:* The keyword from translation module will be taken to form SQL query. Query will be passed to the database for searching and result will be accessed.
6) *Database Module:* Database module will search the database for corresponding English keyword as a attribute value or name. If the entry is found then ontology is build

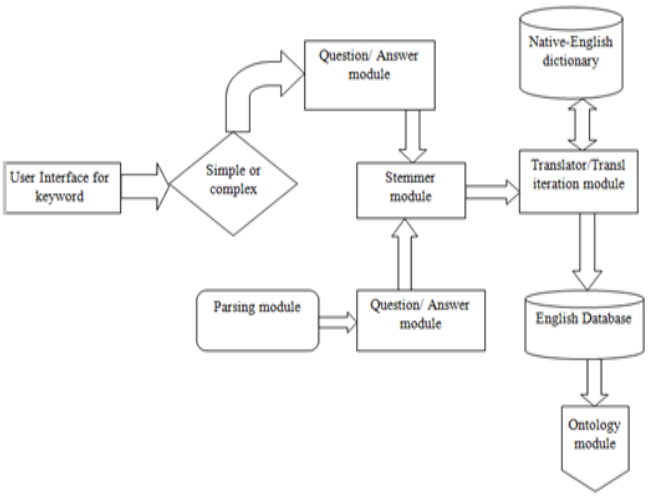

Fig. 1 Flow diagram for ontology development

B. *Proposed Algorithm for Ontology Development*

The proposed algorithm includes following steps *[6]:*

Step1: Enter a keyword generally called proper name for the domain.

Step 2: If keyword is simple, then no parsing is required goto step3 or if it is complex parse the keyword.

Step 3: Check into the database as either table name or attribute name.

Step 4: If it is table name, formulate questions and answers from all the values of its attributes otherwise goto step 5.

Step 5: If it is an attribute or value inside the table, formulate the questions and answers from relevant tuple.

Step 6: Additional questions and answers can be formulated from dependency of the table's attribute.

Step 7: With the answers, every concepts and relationships can be structured to build ontology in a tree or graph like structure. This can be obtained from the databases maintained for the domain.

V. ONTOLOGY BUILDING TOOLS

A broad overview of some available editors and environments that can be used for the building of ontologies is provided. Comparison is done by considering different properties of editors [17], [18].

A. *Protégé*

Protégé is a free, open-source Java-based platform. It supports creation, visualization, and manipulation of ontologies in various representation formats. It can be customized to provide domain-friendly support for creating knowledge models and entering data. Protégé can be extended via a plug-in architecture.

B. *Apollo*

Apollo is user-friendly knowledge application that allows a user to model ontology with basic primitives, such as classes, instances, relation, functions and so on. It can inherit other ontologies and their classes.

C. *Swoop*

Swoop, 2004 is an open-source, Webbased OWL ontology editor and browser. Swoop allows comparison of entities and relationship across various ontologies as well as editing and merging.

D. *TopBraid Composer Free Edition*

TopBraid Composer is modeling tool for creation and maintenance of ontology and comes in three editions: Free Edition (FE), Standard Edition (SE) and Maestro Edition. It provides development of RDF and OWL ontologies.

The Ontology building tool doesn't support Indian languages and requires translation in English language for building ontology. Thus there is no tool for direct ontology building that allows Indian language as an input

Table II represents that most of the work done for building ontology in Indian languages is using a tool. Thus more work has to be done for automatic ontology building.

TABLE I
COMPARISON OF ONTOLOGY BUILDING TOOLS

No	Features	Apollo	Protégé	Swoop	Topbraid composer
1	Availability	Open Source	Open Source	Open Source	Software License
2	Extensibility	Plug-ins	Plug-ins	Plug-ins	Plug-ins
3	Backup Management	No	No	No	Yes
4	Ontology Storage	Files	File and DBMS	HTML Models	DBMS
5	Graphical Taxonomy	No	Yes	Yes	Yes
6	Ontology Libraries	Yes	Yes	No	Yes
7	Semantic Web Architecture	Standalone	Standalone and Eclipse Client/server	Web-based and client-server	Standalone Eclipse Plugins
8	Reasoner	Prolog reasoner	Pellet	Pellet	Pellet
9	Querying	Yes	Yes	No	Yes
10	Indian language support	No	No	No	No

TABLE III
COMPARISON OF DIFFERENT WAYS IN WHICH ONTOLOGY WAS BUILT FOR INDIAN LANGUAGE

Sr no.	Paper	Tool used	Remark
1	Personalization of user profile: creating user profile ontology for **Tamil Nadu** Tourism [5]	Protege	Tool is used for building ontology.
2	Domain specific ontology extractor for **Indian languages** [9]	No tool used, K-partite graph learning algorithm	No NLP tool used. Can be easily adopted for other Indian languages
3	Building **Manipuri-English** machine readable dictionary by implementing ontology [19]	Protege 3.2	Tool is used for building ontology.
4	Domain specific ontology based query processing system for **Urdu language** [20]	Protege	Ontology is constructed using tool and stored in RDF/XML format.

VI. CONCLUSIONS

Currently for building ontologies expansion of Wordnet is necessary. Very little work has been done for developing ontology in Indian languages. The reason for this can be attributed to the fact that number of challenges exists for the construction of ontology for minority languages. Another reason is the lack of knowledge about different Indian languages. There is lot of work related to ontology that will be done in future for different Indian minority languages as the literature or data sets, in these languages will grow and become standardized as well. Future work is required for the improvement in recall and precision of the retrieved documents using ontology. After performing a review on different types of techniques in building ontology it can be concluded that either ontology can be build using CLIR technique or by k-partite graph for the minority languages.

REFERENCES

[1] Staab, S., and Studer, R. Handbook on Ontologies, International handbooks on Information Systems, 2nd ed., London, New York: Springer Dordrecht Heidelberg, 2009.

[2] Kong, H., Hwang, M., and Kim, P. "Design of the automatic ontology building system about the specific domain knowledge," In *Proceedings of 8th International Conference on Advanced Communication Technology, ICACT, IEEE , 2*006, pp. 1405- 1408.

[3] Saraswathi, S., Siddhiqaa, A. M., Kalaimagal, K., and Kalaiyarasi, M. "BiLingual information retrieval system for English and Tamil," *Journal of Computing,* vol. 2, pp. 85-89, April 2010.

[4] Talita, P., Yeo, A. W., and Kulathuramaiyer, N. "Challenges in building domain ontology for minority languages," in *Proc. of IEEE International Conference on Computer Applications and Industrial Electronics, 2010, pp. 574-578.*

[5] Ananthapadmanaban, K. R, and Srivatsa, S. K. "Personalization of user profile: creating user profile ontology for Tamilnadu Tourism," *International Journal of Computer Applications, IJCA,* vol. 23, pp. 42-47, June 2011.

[6] Chaware, S. M., and Rao, S. "Ontology approach for cross language information retrieval," *International Journal of Computer Technology and Applications,* vol. 2, pp. 379-384, 2011. 5. 5.

[7] Durga, A. K., and Govardhan, A. "Ontology based text Categorization- Telgu documents," *International Journal of Scientific & Engineering Research, vol. 2, pp. 1-4 , Sep. 2011.*

[8] Saraswathi, S., Jemibha, P., Sugandhi, M., Mathimozhi, M., Sophia, A. L., and Nagarathinam, A. "Semi-automatic ontology based bilingual information retrieval system (Pilgrimage tourism in South India)," *International Journal of Intelligent Systems and Applications, MECS,* vol. 4, pp. 48-55, April 2012.

[9] Bhatt, B., and Bhattacharyya, P. "Domain specific ontology extractor for Indian languages," In *Proceedings of 10th Workshop on Asian Language Resources, COLING, Mumbai,* 2012, pp. 75-84

[10] Chaware, S., and Rao, S. "Ontology supported inference system for Hindi and Punjabi," In *Proceedings of IEEE International Conference on Technology Enhanced Education,* 2012, pp. 1-6.

[11] Mathur, I., Darbari, H., and Joshi, N. "Domain ontology development for communicable diseases," *CS & IT-CSCP*, vol. 3, pp. 350-360, 2013.

[12] Liu, J. N. K., He, Y., Lin, E. H. Y., and Wang, X. "Domain ontology graph model and its application in Chinese text classification," *Neural Computing & Applications, Springer, London,* vol. 24 , pp. 779-798, March 2014.

[13] Rajveer Kaur, Saurabh Sharma. "Pre-processing of domain ontology graph generation system in Punjabi," *International Journal of Engineering Trends and Technology (IJETT) vol. 17, Number 3- Nov 2014*

[14] Sandeep Chaware, and Srikantha Rao "Integrated approach to ontology development methodology with case study," *International Journal of Database Management Systems (IJDMS) vol. 2, NO. 3 , August 2010*

[15] Oscar Corcho, Mariano Fernandez-Lopez, Asuncion Gomez-Perez "Methodologies, tools and languages for building ontologies. Where is their meeting point?," *www.ComputerScience Web.com*

[16] Fernández López "Overview Of Methodologies For Building Ontologies," *http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-18/*

[17] Sabin Corneliu Buraga, Liliana Cojocaru, Ovidiu Cătălin Nichifor. "Survey on Web Ontology Editing Tools," *Buletinul Stiintific al Universitatii "Politehnica" din Timisoara, ROMANIA, Vol.NN (ZZ), 2006*

[18] Pallavi Grover, Sonal Chawla. "Evaluation of Ontology Creation Tools," *International Journal of Soft Computing and Engineering (IJSCE), Volume-4, Issue-2, May 2014*

[19] Shantikumar Ningombam, Sagolsem, Dr. Bipul Syam "Building Manipuri-English machine readable dictionary by implementing ontology," *International Journal of Engineering Science and Technology vol. 2, NO. 3 , Oct 2011*

[20] Rukhsana Thaker, Ajay Goel "Domain specific ontology based query processing system for Urdu language," *International Journal of Computer Application vol.121, NO.13, July 2015*