Domain Specific Ontology Creation for Marathi Language

Pooja Jidge
Department of Computer Engineering PCE,
Mumbai University, India

Sharvari Govilkar
Department of Computer Engineering PCE,
Mumbai University, India

ABSTRACT
Ontology is one of the central areas in the field of natural language processing (NLP) and artificial intelligence. Today, millions of documents are present in Indian regional languages. To build ontology for these documents manually is time consuming and expensive task. Therefore, this paper presents a system to build an automatic ontology from the unstructured text. Since the unstructured text is widely available still it is difficult for users to find the relation between different concepts. Ontology provides a graphical representation of different concepts that are easily understandable by the user. We have limited our work to Marathi language only. The objective is to propose a system that will accept unstructured Marathi language text as input and provide ontology tree as an output.

Keywords
Ontology, k partite graph, F- measure

1. INTRODUCTION
Over the years, the volume of information available on the World Wide Web is increasing continuously. The rapid development of information technology has led to a collection of large documents in Indian regional languages. To find the relation between different documents manually is a tedious job. Therefore, one possible solution to this problem is automatic ontology building, which finds out the relation between different concepts. Ontology allows to represent domain-specific concepts. Thus, the ontology build can be used for a refined search. By representing a document in knowledge form and giving this as an input base to search engines enables to give more relevant results.

Ontology enables to explain the relation between the concepts and thus better information is served to the user. Information from different sources is considered which represent a variety of heterogeneous data. In ontology, domain specific data are considered and integrated into a single information node. This is presented to the user, taking into consideration the context of the user. This solves the problem of semantic differences.

Marathi is an Indo-Aryan language spoken predominantly by Marathi people of Maharashtra. It is morphological rich and inflectional language. Hence, ontology building of Marathi unstructured text documents is a difficult task. A single root word is formed by combining many morphological variants with inflections, therefore, getting root word from Marathi language is difficult task

Manual construction of ontology is time consuming and costly affair that requires participation of experts having knowledge regarding particular domain. Therefore, automatic ontology building system is required which will automate this process. The proposed system can be useful in moving from a keyword- based approach towards the concept based learning. The idea is to propose a system that will accept Marathi text as input and process the input by building a hierarchical structure of concepts.

2. LITERATURE SURVEY
This section cites the relevant past literature that uses various ontology building techniques. Most of the researchers used cross-language information retrieval technique to convert the document into English language and then build ontology using different tools. The ontology build was again translated into required language using machine translation.

Saraswathi, Asma [1] presents information retrieval system for English and Tamil languages. In this, the author has performed information retrieval on the festival domain. The user can query in any language and the system will provide the output in the proposed language. The authors used ontological tree for inter-language conversion and thus allows the user to query in their native language (i.e. Tamil in this case).

S.M.Chaware, Srikantha Rao, [2] presents an approach for building ontology for grocery shop domain. The author proposed an approach to create ontology from relational databases. The ontology is built dynamically for the grocery shop domain as per the user's requirement. The system architecture includes various different modules such as user interface module, parsing module, Q/A module, stemmer module, translator/transliteration module, query module, database module. The result obtained shows the whole, simple and easy creation of ontology from database.

Panceras Talita, Alvin Yeo, Narayanan Kulathuramaiyer, [3] have proposed challenges and solution for building domain specific ontologies for indigenous languages. The author used Iban as the main language to build ontology for agricultural domain and has presented challenges that arise for building ontology in one domain concept. Several minority languages are morphologically rich languages and lack sufficient resources, expertise and knowledge. Moreover the cost and time constraint makes the ontology building difficult for minority languages.

K.R. Ananthapadmanaban, Dr. S.K. Srivatsa, [4] designed user profile ontology for Tamil Nadu tourism. By identifying the interest of the user, the system suggests appropriate package of tourism for the Tamil Nadu region. The user has to register and fill forms, giving all details and their area of interest. Based on this, tourism ontology inference is made and a perfect holiday destination is generated.

Dr. S. Saraswathi, A. Nagarathinam, [5] designed semi-automatic ontology tree which are created partially manual and it is completed dynamically. The semi- automatic ontology tool improved the efficiency of retrieval of information relating to the users query. The author created

semi-automatic ontology tree for four levels mainly state, district, pilgrimage place and attributes related to pilgrimage place. The XML was used for building ontological tree. The system was tested for 5 different types of input queries and their resultant documents are analyzed using manually created ontology and semi automatic ontology. The system showed improvement in the result when semi-automatic ontology was used in bilingual IR system.

Brijesh Bhatt, Pushpak Bhattacharyya, [6] proposes k-partite graph learning algorithmic program for Indian languages that extract ontology from unstructured data. The result of this approach shows improvement in precision without affecting F-score. This approach reduces complexities for the ontology building.

Sandeep Chaware, Srikantha Rao, [7] have proposed an approach to build ontology from the relational database with some additional rules. The ontology can be built dynamically as per user's need, which will give overall knowledge domain to the user. The system design includes seven modules. The result obtained from the proposed approach will give inference accurately for any type of query.

Rajveer Kaur, Saurabh Sharma, [8] have proposed the pre-processing phase for ontology graph generation in the Punjabi language. The author created Punjabi gazetteer list and stop word list. The list was created manually since no such lexical resources are available online. The pre-processing phase includes removing punctuation, duplicate words, special symbols, and matching the words from dictionary and gazetteer list to the input terms.

The review of existing literature reveals that not much work has been done for automatic ontology building for Marathi script. Ontology building for Indian languages is difficult than English script, as Indian languages are rich in morphemes. Moreover, lack of sufficient resources, expertise and knowledge has made ontology building for Indian languages more difficult. Most of the researchers have used machine translation for ontology building. CLIR/machine translation used for building ontology provides an easy approach for building ontology as it enables us to use the existing resources. However, this technique of CLIR doesn't provide efficient results due to the limitation of machine translation.

3. PROPOSED SYSTEM

The proposed work is to develop domain specific automatic ontology building system for Marathi text which will create hierarchical structure by identifying subsumption relations between the terms. It will use dictionary which makes it possible to share concepts, that is, agreed vocabulary for exchanging information. The input to the system is Marathi text document and output will be ontology tree/concept hierarchy.

4. METHODOLOGY

The system presents a k-partite graph learning algorithm for ontology construction from unstructured text. The algorithm divides the initial set of terms into different partitions based on information content of the terms and then constructs ontology by detecting subsumption relation between terms in different partitions. This approach not only reduces the amount of computation required for ontology construction but also provides an additional level of term filtering. The input to the system is Marathi text document and output will be ontology tree/ concept hierarchy. The detailed architecture of the proposed system is given in figure 1.

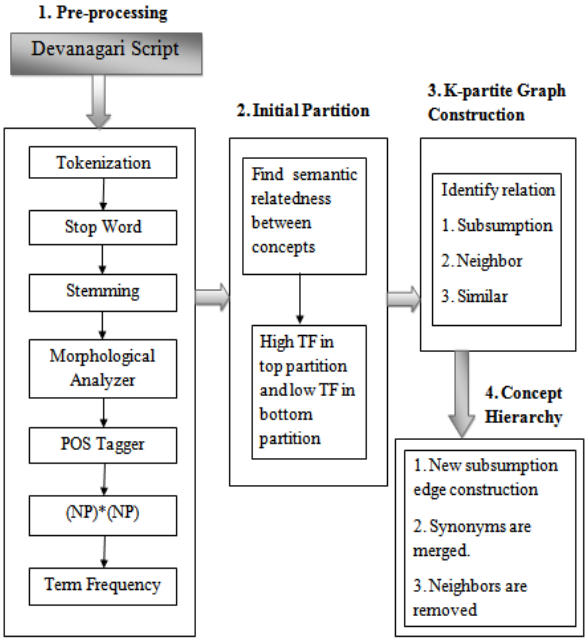

Fig 1: Detailed Architecture of the proposed system

The proposed approach consists of the following module:

1. Preprocessing
2. Initial partition creation
3. K- partite graph construction
4. Concept hierarchy creation

4.1 Preprocessing

4.1.1 Validation of input document

The first step is to preprocess the input document. This module analyzes whether the input document is valid in Devanagari script or not. The words which are not valid according to Devanagari script are simply removed from further processing. Here, Marathi text document is used for validation.

4.1.2 Tokenization

Tokenization is the process of separating the tokens from the input text. This tokenization task is possible by searching spaces between the words. The words are separated from the sentence and treated as a single token.

4.1.3 Stop word removal

Stop words are frequently occurring words in the collection of documents that makes up a large portion of text in document. The stop word is present even after processing. Thus, it is good practice to identify stop words and eliminate them as it tends to take a lot of disk space and reduces overall retrieval performance. Thus it is necessary to remove such stop words from index to save the searching time and enhance searching performance.

A list of all possible stop words occurring in Devanagari script is made and comparison is performed to remove stop words from further processing

4.1.4 Stemmer

Stemming is an important step in the system, which uses a suffix list to remove suffixes from words and thus reduces the word to its stem. The result of stemming is stem of word that

can be given as input to Morphological Analyzer for further processing.

4.1.5 Morphological analyzer

The words after stemming are analyzed to check whether they are inflected or not. The aim of morphological analysis is to recognize the inner structure of the word. A morphological analyzer is expected to produce root words for a given input document. The root and stem of the word may differ in their forms.

4.1.6 POS tag generator

This phase assigns corresponding part of speech tags to the words. A well-chosen tag-set is important to represent parts of speech. The tag-set represents parts of speech.

4.2 Partition

Partition module divides different concept nodes into a different partition. We are considering three levels of partition. Words with highest frequency appear in top partition and lowest frequency word appears in bottom partition. The top partition includes terms having low information content whereas bottom partition consists of terms having detailed information content.

4.3 K- Partite graph construction

K- Partite graph module creates a relation edge between nodes of different partition if there is a semantic relation between the concepts. Wordnet and lexico-syntactic pattern is used to detect the type of relation between semantically related nodes. Three types of relations are considered that mainly includes subsumption, neighbor and similar relation. Wordnet related to a specific domain is build which will be used to detect subsumption and similar relation. Lexico-syntactic pattern will be used to determine neighbor relation.

4.4 Concept hierarchy creation

In this module, the previously created hierarchy is refined. In concept nodes, the nodes that don't have any incoming or outgoing edges are removed. Weakly connected nodes and relation edge that are not representative of the domain are removed. This provides an additional level of term filtering.

5. PSEUDO CODE OF THE PROPOSED SYSTEM

There are different steps used in the proposed system. These steps are given as follows.

Input: Paragraph on specific domain

Output: Hierarchical structure of concepts

Step 1: The paragraph related to specific domain is given as input to the system.

Step 2: Preprocessing the data to obtain root word.

Step 3: POS tagging of the sentence to obtain only noun words.

Step 4: Only NP*NP terms are selected.

Step 5: Semantic relatedness between concepts is measured using similarity measure.

Step 6: Terms having cosine similarity above a threshold value are selected.

Step 7: Terms are arranged into 3 partitions according to term frequency.

Step 8: K- partite module: Form the hierarchy from data to display ontology.

Step 9: Define relation between concepts of ontology using Wordnet and lexico-syntactic pattern.

Step 10: Concept hierarchy module: Remove the weak nodes and relation edges that do not represent the domain.

6. CONCLUSION

Presently there are around 185 tools for ontology building for the English language. However, there is no such tool available for ontology building for Indian regional languages. Very little work has been done for developing ontology in Indian languages. The reason for this can be attributed to the fact that the numbers of challenges for the construction of ontology for Indian languages are many and varied. One of which is the lack of expansion in Wordnet. There are limited numbers of terms in Marathi Wordnet and they generally, do not contain domain specific terms. Another reason is the lack of knowledge, resources, expertise and different morphological and grammatical structure of Indian languages since Indian languages are morphologically rich compared to the English language.

Ontology building using k partite graph technique not only reduces the amount of computation required for ontology construction but also provides an additional level of term filtering. The k-partite graph technique is unsupervised and does not require any human intervention. This approach will build ontology directly for the Marathi language without the need of using machine translation. This will ensure efficient results since the limitation of machine translations are avoided.

7. ACKNOWLEDGMENTS

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. I express my gratitude towards Head of Computer Engineering Department, Dr. Madhumita Chatterjee and to the Principal of Pillai College of Engineering, Dr. R. I. K. Moorthy for extending his support.

8. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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. 36

[5] 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.

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

[7] 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.

[8] Rajveer Kaur, Saurabh Sharma "Pre-processing of Domain Ontology Graph Generation System in Punjabi", International Journal of Engineering Trends and Technology (IJETT) – Volume 17 Number 3 – Nov 2014.

.