

DOMAIN ONTOLOGY AND GRANULAR CONCEPTUALIZATION FOR PUNJABI LANGUAGE

Gursimran Kaur¹, Jaswinder Singh Bhatia²

¹ Assistant Professor, Department of computer science, Chandigarh University,

² Assistant Professor, Department of computer science, Chandigarh University,

Abstract- *Natural Language Processing aims at designing a system which can perform processing like humans. The main concern is to understand how computers can understand and manipulate human language through speech and text. A great deal of work is done in issues and uses of Natural language handling with high precision on foreign dialects however on Indian dialects this is still on beginning stages. We present a framework to create domain ontology for Punjabi language which is express portrayal of ideas in the language. By having domain ontology explicit to the language, we can build up a vocabulary which is allocated with idea chain of command in the domain of this language. We construct a controlled vocabulary of ideas for Punjabi words in form of domain ontology. This created resource can additionally be utilized in various semantic NLP undertakings in Punjabi language.*

Keywords- *Natural Language Processing, Punjabi Language, Domain Ontology, Conceptualization, Vocabulary resource building*

I. INTRODUCTION

Natural Language Processing is a branch of computer science which deals in what capacity would computers is able to be utilized to comprehend and control normal language through content or discourse. It focuses on various tasks like extraction, retrieval, translation and transliteration from unstructured text. The researches utilize the assembled information of human comprehension of dialects to create frameworks that can perform desired undertakings. Language is the most basic method for correspondence. We use it to express our feelings and convey our thoughts. We adapt to new ideas through it normally. The primary goal is to perceive about the comprehension of individuals towards language so it very well may be utilized with various instruments and strategies to influence computers to perform desired assignments. Working on natural language processing, we are concerned about development of models according to processing in such a way as humans do. The main reason for such models is:

- Advancement of tools for language processing
- Gain better comprehension of human communication

A great deal of work is done in issues and utilization of Natural language processing with high precision on foreign languages however on Indian regional languages; this is still on starting stages. Like other Indian dialects, a great deal of work stays to be done in Punjabi language

Punjabi Language is native language of around 122 million people around the world making it 10th most commonly spoken language. Apart from having rich literature history, like other Indian languages technical advancements are recent for it. It doesn't have many explained corpora, rich word references, dictionaries and all around conceptualized domains. As extensive preparing corpora for various types of tasks are additionally not accessible in much plenitude, thus it is difficult to prepare frameworks to do desired undertakings by following machine language approaches.

In this paper, we present a system to develop domain ontology which will give conceptualization of concepts for Punjabi language in granular form. We build up a vocabulary which is elucidated on with areas from Punjabi language. The vocabulary consists for concept hierarchies which will further be helpful in creating a resource. Domain hierarchies are constituted by specific object of knowledge. For example "Cricket" and "Football" are more specific domain than "games".

Domain hierarchies are utilized to form a resource which contains set of words representing concepts. Levels of domains are indicated to frame a controlled hierarchy. In the top level domain, basic concepts will be included and levels below will be conceptualized accordingly focusing on more granularities of concepts. Once the hierarchies are conceptualized, a controlled vocabulary is formed by framing in terms of domain ontology.

II. ONTOLOGIES

Knowledge about a domain of interest can be depicted by Ontologies. By utilization of these, we can enact knowledge in term of concepts and relationships amongst them. This helps in easy perception for humans and proficient transforming for computers. The ontologies are worked regarding encourage reuse of the area learning, to unequivocally depict the space, explicit area to make a typical comprehension of the structure sharable among various partners in the space, delimit the area

learning from operational commitment, and to investigate the space information[2]. Ontology is a portrayal vocabulary which is particular in some specific domain or subject. It is prescribed to begin the development of ontology by characterizing its domain and scope.

For development of ontology, one should answer several questions: [2]

- What is the domain that the ontology will cover?
- For what we are going to use the ontology?
- For what types of questions the information in the ontology should provide answers?
- Who will use and maintain the ontology?

Ontology of domain is usually not final objective. It is identified with characterizing set of words and their hierarchical structure. It is basically conceptualization of domain of specific interest.

III. RELATED WORK

There is no such work as form of domain ontology in Punjabi language and the present work, to the best of our insight, is new in the area. The following section however looks at some of the works of similar nature and furthermore identified with the essential structure of the present work.

Natalya F. Noy and Deborah L. McGuinness[1] have proposed a manual to create ontology. It gives purposes behind which ontology can be created by individuals. The proposed guide gives explicit depiction of idea, properties, attributes, limitations on domains. It describes ontology development methodology for declarative frame based system. The paper has indicated steps for ontology development process. The paper specifies that ontology is a inventive process and no two ontologies can be same which are created by two distinct individuals.

Gobinda G. chowdhury [2] has discussed the fundamentals of NLP and the possible research fields in it. The paper additionally covers vital tasks and uses of NLP including Natural Language Text Processing. He covers the vital issues in NLP. He indicates diverse programming, tools and systems which are utilized for various exercises of NLP.

Rada Milhalcea et.al [3] has worked on utilizing current Natural Language Processing procedures to utilize of common natural languages so that non-expert clients can access more to programming. Her paper indicates that how NLP can be useful to user for understanding and learning of programming language.

Vishal Gupta and G.S. Lehal [4] have talked about condition based Named Entity Recognition (NER) approach for Punjabi content summarization with regards to different NLP undertakings. They have manufactured an asset list which is prior not available. Resources for prefix rule list, addition rule list, center name list, last name list and appropriate name list were designed.

Kamaldeep Kaur and Vishal[5] Gupta have attempted to build a hybrid approach for Punjabi language that characterize words which represent proper names in content into predefined fields. In this paper, first different approaches of Named Entity Recognition (NER) for carefully assembled methodology and machine learning approach are surveyed and a hybrid approach is introduced.

Emilia Stoica and Marti A. Hearst [6] have exhibited an automated methodology for forming metadata hierarchies. In this, the WordNet Hierarchy is transformed to reflect content of target data accumulation. They have explored approximately 35000 art documents containing about 23000 unique words. An algorithm is also used to compare the results of this algorithm with WordNet results.

Satoshi Sekine, Kiyoshi Sudo and Chikashi Nobata [7] have displayed a design of hierarchy which contains more Named Entity types. This paper proposes a Named Entity hierarchy which contains about 150 NE types. They indicate three phases of development of hierarchy. First is building initial hierarchy, merging hierarchies, refining hierarchies. This resource can be utilized in many NLP applications like information retrieval, machine translation and resource for other extended hierarchies.

Sujan Kumar Saha et.al [8] has portrayed a hybrid framework that is utilized for Named Entity Recognition (NER) of different Indian languages. It applies Maximum Entropy model, language explicit tenets to NER. Initially a pattern is manufactured including corpora and language explicit features. The above model is tried on Hindi, Bengali, Oriya, Telgu and Urdu.

Kavi Mahesh and Sergei Nirenburg [9] portray that each NLP framework which try to represent or control text need a ontology which gives classification of ideas which can further be utilized as semantic natives. The fundamental objective of paper is to build up a framework that can create thorough text significance to enter message in any set of source language. The framework created will be proficient in taking care of issues related to semantics in NLP.

Hammad Afzal, Robert Steves and Goran Nenadic [10] depict a system which constructs a controlled vocabulary to portray bio informatics services. In this a methodology is utilized that blend lexical and relevant profiles of applicant terms to propose terms in vocabulary.

IV. METHODOLOGY

We built up the domain ontology by allotting ideas as top dimension domains and after that populate lower level domains having progressively granular conceptualization under higher level domains. By this a controlled vocabulary with explicitness to Punjabi language is being sorted out. We use Jena API [11] in Java to build up a ontology delivered in Web

Ontology Language (OWL) [12]. OWL is a Semantic Web language which is intended to learning about things, groups and relationships between them. After the development of ontology we do assess the ontology by manual ways to check whether it is filling the required need or not.

Ontologies are comprised of classes organized into hierarchies, relationships and individuals to represent knowledge about explicit domain.

Classes: Classes in the ontologies represent the principle ideas in the field of interest . For instance, "Education and Science " (ਸਿੱਖਿਆ ਅਤੇ ਵਿਗਿਆਨ) class catches different words of Punjabi that represent different features identified related to given class. There can be further subclasses dependent on sub concepts under the principle classes.

For instance, subclass under the Education and science (ਸਿੱਖਿਆ ਅਤੇ ਵਿਗਿਆਨ)" class is subclass "Engineering (ਇੰਜੀਨੀਅਰਿੰਗ)". The classes are organised in hierarchy according to relation to specificity as indicated by a language or subject.

Individuals: The elements in ontology are enacted by individuals. For example: “computer science engineering (ਕੰਪਿਊਟਰ ਸਾਇੰਸ ਇੰਜੀਨੀਅਰਿੰਗ)” is an element or individual under the subclass “Engineering (ਇੰਜੀਨੀਅਰਿੰਗ)”.

Top-Level Domains: For creating class hierarchy systems, we utilize top-down methodology. We begin with most general ideas first and after that include values to the consequent fields. Concepts are recognized as a matter of first importance to be considered as top dimension areas to cover the words in vocabulary. Following Fig 1 shows some of the top level domains.



Figure 1 Top Level Domains

Sub Domains: Currently we have focused granularity of words up to three levels of domains which can further be classified into finer domains in the generation of future resource.

According to the top domains we extend sub domains as shown in figure 2

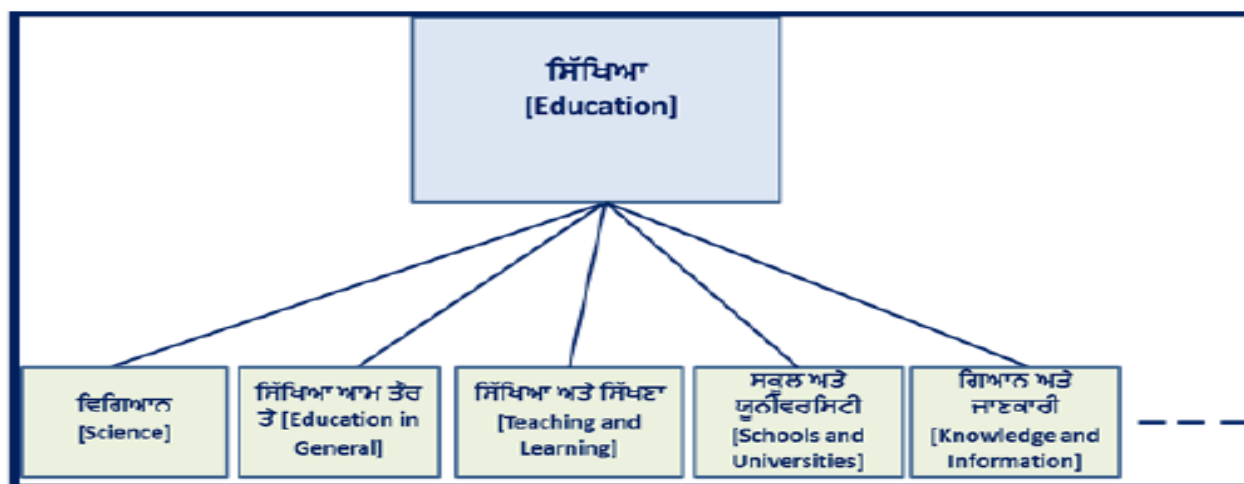


Figure 2 Sub Domain for Education (ਸਿੱਖਿਆ) top level domain

Eventually we will make a corpus with explicitness to Punjabi language. It can be used as a semantic resource to the other technical developments in the field of NLP with respect to the Punjabi language in near future.

V. EVALUATION

Ontology development requires evaluation of results about ontology with the goal so that we can confirm that whether it serves its motivation effectively or not. Different methodologies for assessment can be utilized. These approaches include comparison with sources of data for ontology and evaluation to satisfy set of predefined requirement by human to access the ontology. The ontology is to be assessed by its consistency, precision and utility. Following are the basis on which evaluation is finished:

- **Evaluation of consistency:** It is done based on meaning of class, its properties and constraints. We can check whether it is practical for class to have occurrences or not. A class that can't have instances will be inconsistent.
- **Evaluation of precision:** In it, the precision is checked by cross validating qualities with original sources. It is done manually. We can likewise utilize an inquiry based mechanized way to deal with check precision.
- **Evaluation of utility:** It tends to be done by utilizing the knowledge learnt by ontology in various applications. The information can likewise be checked physically with the goal that both the outcomes can be compared.

VI. CONCLUSION AND FUTURE WORK

Resources are of great significance in Natural Language Processing. These can range from being basic and general to sophisticated and concentrated. A resource can be worked by focusing multiple NLP tasks. A resource can be in multiple forms such as references, corpora, dictionaries, annotated corpora and so on. They give better comprehension of common language levels. These dimensions are lexical, morphological, syntactic, and semantic discourse levels.

We have completed the conceptualization and ontology populace of two levels of domains. Our work of conceptualizing the domains of third dimension granularity and furthermore allocation the expressions of word reference with these dimensions is under advancement, henceforth the concentration later on will be to finished the assignment soon and after due assessment, spread the information captured in the ontology for more extensive open use through an application or ontology program.

REFERENCES

- [1] Noy, N. F., &McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.
- [2] Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1), 51-89.
- [3] Mihalcea, R., Liu, H., & Lieberman, H. (2006). Nlp (natural language processing) for nlp (natural language programming).In *Computational Linguistics and Intelligent Text Processing* (pp. 319-330). Springer Berlin Heidelberg.
- [4] Gupta, V., &Lehal, G. S. (2011). Named Entity Recognition for Punjabi Language Text Summarization. *International Journal of Computer Applications*, 33(3), 28-32.
- [5] Kaur, K., & Gupta, V. (2012). Name and Entity Recognition for Punjabi Language. *Machine translation*, 2(3).
- [6] Stoica, E., & Hearst, M. A. (2004, May). Nearly-automated metadata hierarchy creation. In *Proceedings of HLT-NAACL 2004: Short Papers*(pp. 117-120). Association for Computational Linguistics.
- [7] Sekine, S., Sudo, K., &Nobata, C. (2002, May). Extended Named Entity Hierarchy. In *LREC*.
- [8] Saha, S. K., Chatterji, S., Dandapat, S., Sarkar, S., &Mitra, P.(2008, January). A hybrid approach for named entity recognition in indian languages. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian Languages*(pp. 17-24).
- [9] Mahesh, K., &Nirenburg, S. (1995, October). Semantic classification for practical natural language processing. In *Proceedings of Sixth ASIS SIG/CR Classification Research Workshop: An Interdisciplinary Meeting, Chicago IL*.
- [10] Afzal, H., Stevens, R., &Nenadic, G. (2008). Towards semantic annotation of bioinformatics services: building a controlled vocabulary. In *Proc. of the Third International Symposium on Semantic Mining in Biomedicine* (pp. 5-12).
- [11] Jena, java based application programming interface (API). [8 March 2016, Date last accessed]; Available from: <http://jena.sourceforge.net/>.
- [12] Web Ontology Language (OWL). [8 March 2016, Date last accessed]; Available from: <http://www.w3.org/TR/owl-ref/>.