Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 3, Issue 2; January-March, 2016, pp. 119-122

http://www.krishisanskriti.org/Publication.html

Explicit Conceptualization and Ontological Representation of Domains for Punjabi Words

Gurinder Pal Singh Gosal[1], Gursimran Kaur[2] and Neeraj Sharma[3]

[1,2,3]Department of Computer Science, Punjabi University, Patiala
E-mail: [1]gurinder@pbi.ac.in, [2]gursimran.kaur6499@gmail.com, [3]sharma_neeraj@pbi.ac.in

Abstract—*Resources are of great importance in Natural Language Processing (NLP) and NLP resources are usually built by targeting multiple NLP tasks. It has been observed commonly that many of NLP resources end up not having higher utility and acceptance amongst the targeted user community even if a lot of time and energy is used in building them. Like many other Indian languages, Punjabi language also inherits a rich literature history but technical developments are very recent for it. It does not have many NLP resources of its own, such as, annotated corpora, rich dictionaries, sentiment lexicons, well conceptualized domains etc. Our present work is an effort to build a controlled vocabulary of concepts or domains for Punjabi words in the form of domain ontology. Ontologies capture and describe the current state of knowledge about a domain of interest, and represent it in terms of concepts and relationships in ways that computers can process efficiently and humans can understand easily. We develop the domain ontology by starting assigning concepts as top level domains and then populate lower level domains having more granular conceptualization under the higher level domains. This developed resource can further be used in different semantically based NLP tasks in Punjabi language.*

1. INTRODUCTION

Punjabi is 10th most widely spoken language in the world. Punjabi language is native language of more than 130 million people. Not only in Punjab and northern parts of India, it is also spoken in Pakistan and other parts of world including America, Canada and Europe etc. Like other Indian languages, Punjabi inherits rich literature history but technical developments are new for it. Like other Indian languages, a lot of work remains to be done in the field of Punjabi language. It also does not have many resources of NLP like annotated corpora, rich dictionaries, lexicons and well conceptualized domains.

In this paper, we present a system to develop domain ontology which will be explicit representation of concepts for words in Punjabi language. We develop a vocabulary which is annotated with domains from within this concept hierarchy in Punjabi language. Domain hierarchies are represented by specific object of knowledge. For example, "*Fish*" or "*Birds*" are more specific domains than "*Animals*". Domain hierarchies are used to create a resource which contains set of words representing concepts. Levels of domains are specified to form 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 representing it in terms of domain ontology.

2. ONTOLOGIES

Knowledge about a domain of interest can be described by Ontologies. By use of Ontologies we can represent knowledge in term of concepts and relationships amongst these concepts. This helps in easy understanding for humans and efficient processing for computers. Thomas Gruber gives a simple definition of ontology as "an explicit specification of a conceptualization" [1]. The ontologies are built with respect to facilitate reuse of the domain knowledge, to explicitly describe the domain, specific domain to make a common understanding of the structure sharable among different stakeholders in the domain, delimit the domain knowledge from operational obligation, and to explore the domain knowledge [2].

Ontology is a representation vocabulary which is specialized in some specific domain or subject. It is recommended to start the development of ontology by defining its domain and scope.

That is, one should answer several basic 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?

In practical terms, developing ontology includes [2]:

- Defining classes in the ontology
- Arranging the classes in a subclass/super class hierarchy.

- Defining slots and describing allowed values for these slots.
- Filling in the values for slots for instances

3. RELATED WORKS

There is no such work as kind of domain ontology in Punjabi language and the current work, to the best of our knowledge, is new in the area. The following section however looks at some of the works of similar nature and also related to the foundational framework of the present work.

Gobinda G. chowdhury [3] has talked about the basics of NLP and the possible research areas in it. The paper also covers important tasks and applications of NLP including Natural Language Text Processing. He covers the major issues in NLP. He specifies different software, tools and techniques which are used for different activities of NLP. Natalya F. Noy and Deborah L. McGuinness [2] have proposed a guide to create ontology. It gives reasons for which ontology can be developed by people. The proposed guide gives explicit description of concept, properties, attributes, restrictions on domains. It describes ontology development methodology for declarative frame based system. The paper has listed steps for ontology development process. The paper specifies that ontology is a creative process and no two ontologies can be same which are developed by two different people.

Rada Milhalcea et.al [4] has worked on using modern Natural Language Processing techniques to make use of natural languages so that non-expert users can access more to programming. Her paper specifies that how NLP can be helpful to user for understanding and learning of programming language. Elizabeth D. Liddy [5] has definition of NLP which covers the aspects that are part of other definitions available. She has referred to various goals and applications of NLP. She also talks about different levels in NLP. Different approaches for NLP are also explained along with their comparisons on different aspects.

Vishal Gupta and G.S. Lehal [6] have discussed condition based Named Entity Recognition (NER) approach for Punjabi text summarization in the context of various NLP tasks. They have built a resource list which is earlier not available. Resources for prefix rule list, suffix rule list, middle name list, last name list and proper name list were formed. Kamaldeep Kaur and Vishal Gupta [7] have tried to build a hybrid approach for Punjabi language that classify words which represent proper names in text into predefined domains. In this paper, first various approaches of Named Entity Recognition (NER) for handcrafted approach and machine learning approach are surveyed and a hybrid approach is presented.

Emilia Stoica and Marti A. Hearst [8] have presented an automated approach for creating metadata hierarchies. In this, the WordNet Hierarchy is converted to reflect content of target information collection. They have experimented with 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 [9] have presented a design of hierarchy which contains more Named Entity types. This paper proposes a Named Entity hierarchy which contains about 150 NE types. They specify three stages of development of hierarchy. First is building initial hierarchy, merging hierarchies, refining hierarchies. This resource can be used in many NLP applications like information retrieval, machine translation and resource for other extended hierarchies.

Sujan Kumar Saha et.al [10] has described a hybrid system that is used for Named Entity Recognition (NER) of various Indian languages. It applies Maximum Entropy model, language specific rules to NER. First a baseline is built including corpora and language specific features. The above model is tested on Hindi, Bengali, Oriya, Telgu and Urdu. B. Chandrasekaran, John R. Josephson and V. Richard Benjamins [11] give introduction to ontologies and their role in information system. This paper surveys the developments in field of ontologies. It specifies that ontologies are content theories about the objects. It provides potential terms for describing knowledge.

Kavi Mahesh and Sergei Nirenburg [12] describe that every NLP system which seek to represent or manipulate text need an ontology which gives classification of concepts which can further be used as semantic primitives. The main goal of paper is to develop a system that can produce comprehensive text meaning to input text in any set of source language. The system developed will be efficient in solving problems related to semantics in NLP. John A. Bateman [13] discusses the idea of functions adopted in NLP which are to be fulfilled by ontologies. This paper explicitly starts discussion concerning design and construction of ontology. It gives further explanation to use of ontologies for text generation.

Hammad Afzal, Robert Steves and Goran Nenadic [14] describe a system which builds a controlled vocabulary to describe bio informatics services. In this a methodology is used that combine lexical and contextual profiles of candidate terms to suggest terms in vocabulary. Gurinder Pal Singh Gosal [15] has talked about different methodologies for developing of ontology. It gives an integrated view for different available methodologies in ontology development to look at different activities performed during development process.

4. ONTOLOGY FOR DOMAINS IN PUNJABI

Our domain ontology is developed by starting assigning concepts as top level domains and then populating lower level domains having granular conceptualization under top level domains. We are using top down approach for it. By this a controlled vocabulary with specificity to Punjabi language is being organized. We use Jena API [16] in Java to develop a ontology produced in Web Ontology Language (OWL) [17]. OWL is a Semantic Web language which is designed to

represent knowledge about things, groups and relationships between them. After the development of ontology we do evaluate the ontology by manual means to check whether it is serving the prescribed purpose or not.

Ontologies are made up of classes organized into hierarchies, relationships and individuals to represent knowledge about specific domain.

Classes: Classes in the ontologies represent the main concepts in the field of interest. For example, "*The Law and Crime (ਕਾਨੂੰਨ ਅਤੇ ਅਪਰਾਧ)*" class captures various words of Punjabi that represent various aspects or features related to crime and law.

There can be further **subclasses** based on sub concepts under the main classes. For example, subclass under the *The Law and Crime (ਕਾਨੂੰਨ ਅਤੇ ਅਪਰਾਧ)*" class is subclass "*Punishment (ਸਜ਼ਾ)*".

The classes are organized in hierarchies based on relation to specificity according to a language or subject.

Individuals: The elements or entries in ontology are represented by individuals. For example: "*capital punishment (ਫਾਂਸੀ)*" is an element or individual under the subclass "*Punishment (ਸਜ਼ਾ)*".

Top-Level Domains

For developing class hierarchies, we use top-down approach. We start with most general concepts first and then add values to the subsequent domains. Concepts are identified first of all to be considered as top level domains to cover the words in vocabulary. Some of the top level domains are shown in *Fig. 1*.

Fig. 1: Some Top Level Domains

Sub Domains

Presently we have focused our granularity up to three levels of domains which can further be classified into finer domains in the future.

To the top-level domains we add subdomains as shown in *Figures 2 and 3*. The *Fig. 2* here represents the sub-domains for top level domain "*Work, Business, Industry and Technology (ਕੰਮ-ਕਾਰੋਬਾਰ-ਉਦਯੋਗ-ਤਕਨਾਲੋਜੀ)* domain. We have added some subdomains for it.

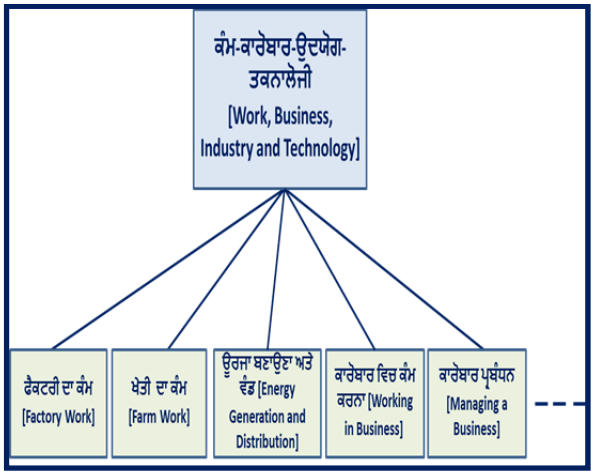

Fig. 2. Sub-Domains for Work, Business, Industry and Technology (ਕੰਮ-ਕਾਰੋਬਾਰ-ਉਦਯੋਗ-ਤਕਨਾਲੋਜੀ) Top-Level Domain

Similarly in *Fig. 3*, we represent some sub-domains for top level domain "*Education (ਸਿੱਖਿਆ)*".

Eventually we will create a corpus with specificity to Punjabi language. It can further be used as a semantic resource to the other technical developments in the field of NLP with respect to the Punjabi language.

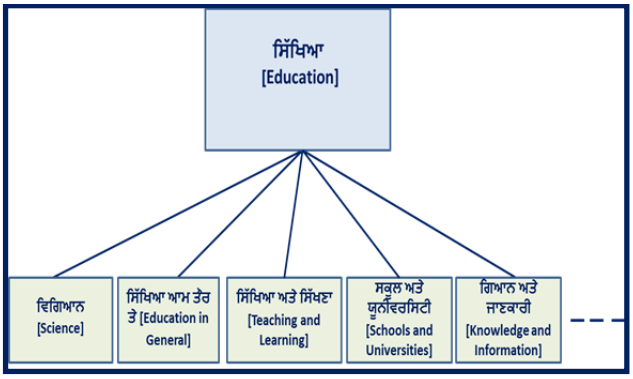

Fig. 3. Sub-Domains for *Education (ਸਿੱਖਿਆ)* Top-Level Domain

5. EVALUATION

Ontology development requires evaluation of resulting ontology so that we can determine whether it does serve its purpose or not. Various approaches for evaluation can be used. 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 evaluated by its consistency, accuracy and utility.

Following are the basis on which evaluation is done:

- **Evaluation of Consistency:** It is done on the basis of definition of class, its properties and constraints. We can check whether it is feasible for class to have instances or not. A class that cannot have instances will be inconsistent.
- **Evaluation of Accuracy:** In it, the accuracy is verified by cross validating values with original sources. It is done manually. We can also use a query based automated approach to check accuracy.
- **Evaluation of Utility:** It can be done by using the knowledge gained by ontology in different applications. The knowledge can also be checked manually so that both the results can be compared.

6. CONCLUSION AND FUTURE WORK

Resources are of great importance in NLP. These can range from being simple and general to sophisticate and specialized. A resource can be built by targeting multiple NLP tasks and it can be in multiple forms such as vocabulary, corpora, dictionaries, lexicons, annotated corpora and so on. Resources give better understanding of natural language levels, such as, lexical, morphological, syntactic, semantic and discourse levels.

Different authors have suggested that while building a particular resource we should follow the established standards so that it can be widely accepted, usable, reliable and durable. With the growing publicity, natural language processing community has set particular standards for resources.

We have finished the conceptualization and ontology population of two levels of domains. Our work of conceptualizing the domains of third level granularity and also assigning the words of dictionary with these levels is under progress, hence the focus in the future will be to complete the task soon and after due evaluation, disseminate the knowledge captured in the ontology for wider public use through an application or ontology browser.

REFERENCES

[1] Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing?. *International journal of human-computer studies*, *43*(5), 907-928.

[2] Noy, N. F., &McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.

[3] Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, *37*(1), 51-89.

[4] 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.

[5] Liddy, E. D. (2001). In Encyclopedia of Library and Information Science, Marcel Decker. *Inc.- Natural Language Processing*.

[6] Gupta, V., &Lehal, G. S. (2011). Named Entity Recognition for Punjabi Language Text Summarization. *International Journal of Computer Applications*, *33*(3), 28-32.

[7] Kaur, K., & Gupta, V. (2012). Name and Entity Recognition for Punjabi Language. *Machine translation*, *2*(3).

[8] 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.

[9] Sekine, S., Sudo, K., &Nobata, C. (2002, May). Extended Named Entity Hierarchy. In *LREC*.

[10] 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).

[11] Chandrasekaran, B., Josephson, J. R., &Benjamins, V. R. (1999). What are ontologies, and why do we need them?. *IEEE Intelligent systems*, (1), 20-26.

[12] 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*.

[13] Bateman, J. A. (1997). The theoretical status of ontologies in natural language processing. *arXiv preprint cmp-lg/9704010*.

[14] 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).

[15] Gosal, G. P. S. Ontology Building: An Integrative View of Methodologies.

[16] Jena, java based application programming interface (API). [8 March 2016, Date last accessed]; Available from: http://jena.sourceforge.net/.

[17] Web Ontology Language (OWL). [8 March 2016, Date last accessed]; Available from: http://www.w3.org/TR/owl-ref/.