PuDO: Development of an Ontology for Hierarchical Organization and Representation of Domains for Punjabi Words

Gurinder Pal Singh Gosal
Department of Computer Science
Punjabi University
Patiala, India
gosal.gps@gmail.com

Neeraj Sharma
Department of Computer Science
Punjabi University
Patiala, India
sharma_neeraj@hotmail.com

Abstract— (Natural Language Processing (NLP) is the area of research which focuses on the different tasks of understanding, extraction and retrieval from unstructured text. It makes use of multiple tools, resources and methodologies for performing these tasks. NLP applications developed depend heavily upon resources apart from tools and methodologies. Like many other Indian languages, Punjabi language also inherits a rich literature history but on technological aspects it is relatively under resourced and still a lot of work remains to be done in the field of Punjabi language processing. There are many researchers, groups and organizations which are working on the different aspects of Punjabi language processing but it does not have many NLP resources of its own, such as, annotated corpora, rich dictionaries, sentiment lexicons, conceptualized domains etc.

Our present work is an attempt to develop a controlled vocabulary of concepts or topics (domains) for Punjabi words and present it in the form of 'domains ontology', **PUDO (Punjabi Domains 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. This paper presents our work which is based on identifying the concepts termed as domains for Punjabi language words which can be organized in a hierarchical manner. The hierarchy is based on relation of specificity for Punjabi language. We developed the domains ontology by starting assigning concepts as top level domains and then conceptualized lower level domains having more granular conceptualization under the higher level domains. The developed ontology is further populated with the words as instances which evoke these domains. This developed resource can further be used in different semantically based NLP tasks in Punjabi language.

Keywords— *Domains, Ontology, Semantics, Natural Language Processing, NLP, Punjabi NLP*

I. INTRODUCTION

Natural Language Processing uses multiple knowledge resources, different methodologies and variety of tools with the aim of understanding, extraction and retrieval from unstructured text. Elizabeth D. Liddy has given a very detailed definition of NLP as "*a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications*" [1].

When we contemplate over resources for NLP, the natural language text itself can be considered as a primary resource for performing multiple NLP tasks and building applications thereupon. However, the production and availability of range of other NLP resources is always desirable as these resources are immensely beneficial to the NLP community at large. These resources have a huge importance of their own and they can range from being very simple and less organized to structured and extremely sophisticated resources. Further these resources can be of very general purpose or can be focussed and specialized one. These resources might be conceived, built and operated by different individual researchers, consortiums or organizations and can be publically available for free or commercially controlled.

The conception and development of NLP resources mostly revolves around multiple fundamental NLP tasks such as, for part-of-speech (POS) assignments, stemmers, lemmatizers, chunkers, syntactic analysers and parsers. However, these are being built also by focussing on sophisticated tasks such as, ontological knowledge representation, summarization, question answering, literature-based discovery, relation and event extractions etc. The development of these resources extends to all levels of understanding natural languages, such as, lexical, morphological, syntactic, semantics and discourse levels. The resources are leveraged in multiple forms, such as, controlled vocabulary terms, corpora, annotated corpora, knowledge-based repositories, dictionaries, lexicons, manually annotated collections, semantic tools, taxonomies, ontologies and so on. These resources are being built by researchers and developers at individual levels, in form of collaborative groups and consortiums, at research organization levels, and with all sorts of financial models like open access, commercial, privately owned or government funded.

As the NLP resources are being built at a very rapid rate by focussing on all the levels of natural language understanding, not all the resources end up being useful and acceptable to the targeted community of users, as has been

ISSN: 2229-6913 (Print), ISSN: 2320-0332 (Online)

observed by Sergei et al. [2]. In their paper while looking at the rationale of building NLP resources, they argued that though a lot of time and energy are being put on while building resources for NLP, it is not always the case that the developed resources are of higher quality, utility and henceforth best ones.

Furthermore, many NLP Resources are usually centered on a particular language and specific domain while other problem being that they are not always available to the whole community. With the growth of publicly available NLP resources with every passing day, the NLP community has started looking for common interchangeable formats, guidelines, and standards for these resources, as the resource building is considered to be expensive and time-consuming. *For example*, if an ontology is being built representing a specific domain of interest, it should follow the established standards of ontology development and should be widely acceptable to the community for establishing itself as a successful resource [3].

There were not many NLP resources available for Indian languages until some years back. However, there is an upsurge in developing different resources for processing of national language of India, Hindi, as well as for many other regional languages. Punjabi language inherits a very rich history from literature and cultural perspective yet it is only in the recent pasts efforts have been initiated for the technical developments of its processing. There have been efforts by many researchers, groups and organizations to work on the different aspects of Punjabi language processing but like other Indian languages, a lot of work remains to be done in the field of Punjabi language processing. It lacks in richness of NLP resources and has limited resources like rich dictionaries, lexicons, annotated corpora, well conceptualized domains etc.

Coming from the category of Indo-Aryan languages and a native language of more than 130 million, Punjabi is tenth most widely spoken language in the world [4]. The people speaking Punjabi are spread all across the globe. Mainly spoken in the state of Punjab and also in northern parts of India, it is also spoken in the Punjab province of Pakistan. The other regions of world including America, Canada and Europe also have a huge population of Punjabi speaking people.

Punjabi language is written in two scripts namely Gurmukhi and Shahmukhi. Gurmukhi script is the script used for writing Punjabi mostly in East Punjab. Punjabi is most commonly written in the Gurmukhi script. Gurmukhi script is considered to be complete and accurate way to represent Punjabi sounds and it follows a 'one sound-one symbol' principle [5]. Shahmukhi script is the way of writing Punjabi language in the Persian script and followed mostly in western Punjab. The present work of conceptualizing and representing domains pertains to Gurmukhi script. The application of the current work to Shahmukhi script needs to be investigated and carefully examined before coming to any conclusion.

Our present work of building a resource of domains for Punjabi language is accomplished by starting assigning concepts as top level domains and then conceptualizing lower level domains having more granular conceptualization under these top level domains. The approach used is motivated by lack of any hierarchy in whatever other little efforts done for available for defining domains. *For example*, in one of such works, "Verbs" and "Adjectives" are represented as main topics but "Grammar" and "Language" which clearly are higher in hierarchy of conceptualization to "Verbs" and "Adjectives" are not even figuring in the topic lists. In our work, the careful consideration to these details is given to have a controlled vocabulary with specificity to concepts in Punjabi language.

Ontologies have become a key tool of conceptualization, data integration and knowledge representation in different domains of interest and these are being built and used in multiple disciplines now a day [6]. The ontologies are seen as a tool for explicit conceptualization of any domain of interest and are understandable to humans and easier for computers to process and manipulate. In our work, the Domains Ontology PUDO (Punjabi Domains Ontology) is developed as a resource, which we use for conceptualization and representation of controlled hierarchy of domains for Punjabi language. The PUDO ontology builds a vocabulary of Punjabi words which have been assigned domains from within the domains hierarchy in the ontology. This hierarchy of domains has been motivated, designed and adapted based on multiple sources available online and worked upon manually under the guidance of language experts.

Domains represented as such in terms of concept hierarchies and in the form of an ontology presents a semantic resource which contains set of words associated with domains of finer granularity.

II. ONTOLOGIEES

Ontologies support integrating knowledge about a particular area of interest from disparate and heterogeneous sources and representing it in a way easily understandable for humans and efficient for processing by computers. Thomas Gruber gave a very simple definition of ontology and termed it as "*an explicit specification of a conceptualization*" [7]. Therefore, in simple words we can represent knowledge of an area of interest in term of some concepts and relationships amongst these concepts by using an ontology. The ontologies apart from explicitly describing the domain of interest, also facilitate reuse of the domain knowledge, make a common understanding of the structure sharable among different stakeholders in the domain, separate the domain knowledge from operational obligation, and give a platform to explore the domain knowledge [8]. Ontologies can be distinguished from relational databases as they enable the integration, mining, and reasoning over diverse data sets by conceptually representing knowledge [9,10].

For the development of ontology, different researchers have envisaged a certain set of activities to be performed in each stage of the ontology development process and different methodologies have been proposed by researchers for formalizing the different stages. Natalya F. Noy and Deborah L. McGuinness [8] in their paper have proposed a guide to create ontology by listing steps for

ISSN: 2229-6913 (Print), ISSN: 2320-0332 (Online)

ontology development process. They describe the reasons for which ontology can be developed by people and give explicit description of concept, properties, attributes, restrictions on domains. In simple practical terms, developing ontology includes: 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 [8]. Gurinder Gosal [11] has talked about different methodologies for developing of ontology. The paper gives an integrated view for different available methodologies in ontology development to look at different activities performed during development process. The paper also influence for choice of approaches for different ontology tasks.

As far as some earlier efforts in the direction of domain conceptualization are concerned, Kamaldeep Kaur and Vishal Gupta [12] have tried to build a hybrid approach for Punjabi language that classify words which represent proper names in text into predefined domains. The Advanced Centre for Technical Development of Punjabi Language Literature and Culture, Punjabi University [13] have developed a topic dictionary organized into more than eighty categories, such as, adjectives, nouns, food, fruits, animals, months etc. categories. This dictionary has around 3100 entries divided into these categories having also the words associated with a picture along with the pronunciation in Punjabi. However, this work is focused more on specialization of some concepts around these words rather than having generalized concepts and no hierarchy in the conceptualization of topics is present there also.

To the best of our knowledge, there is no such reported work of representing domains in terms of an ontology for Punjabi language and the present effort of ours is new in the area.

III. DOMAINS ONTOLOGY

Domains ontology for Punjabi words, PUDO, is developed by starting assigning concepts as top level domains and then having lower level domains with more granular conceptualization under these top level domains. Ontologies are structured by having concepts as classes organized into controlled hierarchies, relationships amongst classes and members or individuals to represent the knowledge weaved around a specific domain of interest.

Classes: The main concepts in the domain of interest are represented by classes in the ontologies. There can be further **subclasses** based on sub concepts under the main classes. The classes are organized in hierarchies based on relation to specificity according to a language or subject. *For example*, "*ਰੋਜ਼ਾਨਾ ਜ਼ਿੰਦਗੀ (Everyday Life)"* class captures various words of Punjabi that represent various aspects or features related to everyday life. Further, a class under the "*ਰੋਜ਼ਾਨਾ ਜ਼ਿੰਦਗੀ (Everyday Life)"* class is subclass *"ਘਰ ਦੀ ਸਫ਼ਾਈ (Cleaning the House)"*.

Relations: The concepts and their hierarchies in ontologies are not the only things needed to describe the underlying domain completely. To further define the internal organization of concepts with various features and characteristics of the concepts, relationships or properties are introduced in ontologies. These properties, used for further characterization of the instances of concepts or showing the connections between instances of various concepts, are termed as data properties or object properties. *For example*, the *Domain* class can have a property *hasDefinition* to show that a particular domain has a definition of the domain occurrence such as "*ਘਰ ਦੀ ਸਫ਼ਾਈ (Cleaning the House)*" and similarly instances of the class *Word* can have a property *hasDomain* to show relationship with instances of other class *Domain*.

Individuals: The elements or entries in ontology are represented by instances or individuals. *For example*: "*ਕਲਾਕਾਰ (Actor)*"is an element or individual or instance under the domain *"ਅਦਾਕਾਰੀ (Acting)"*.

A. PUDO Conceptualization and Population

The PUDO ontology development is centered on the specification of concepts as domains in Punjabi language and then further developing the system of generating the schema and populating the data in the ontology. There are multiple tools and resources available for the development of ontologies and for our domains ontology, we have used Jena API [14] in Java which produces PUDO in a Semantic Web language designed to represent knowledge about concepts and relationships between them and known as Web Ontology Language (OWL) [15].

The following discussion pertains to identifying the concepts in hierarchy of domains:

B. Identification of Concepts as Domains

Starting with the most general concepts at the top, the hierarchy develops in a top-down manner and conceptualizes sub domains at lower levels focusing on finer granularities of concepts.

Top-Level Domains

As the top level domains, we focus in the start on most general concepts of representation and store them under the "*TopLevelDomain*" class under superclass of "*Domain*" in ontology. ***Figure1*** shows some concepts which are identified to be considered as top level domains to cover the words in vocabulary. *For example*, the concept "*ਭਾਸ਼ਾ-ਸੰਚਾਰ (Language-Communication)"* represented as top class encapsulates different words of Punjabi which characterize different aspects or features related to 'language and communication' domain.

ISSN: 2229-6913 (Print), ISSN: 2320-0332 (Online)

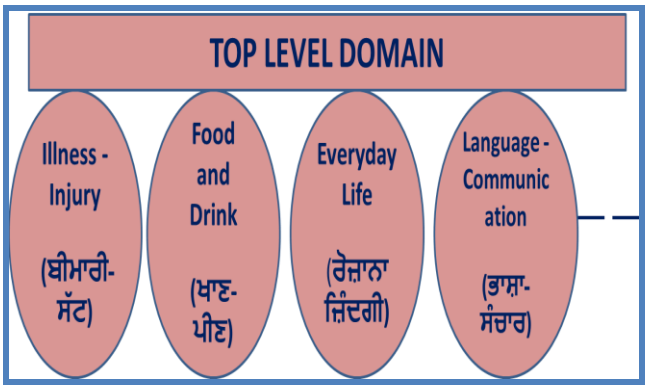

Figure 1: Some Top Level Domains in Punjabi Domains Ontology

Sub-Domains

For developing class hierarchies, we add concepts to serve as sub-domains under the higher level domains. Presently we have focused our granularity up to three levels of domains which can further be classified into finer domains as the need may be in the future

To the top-level domains, we add subdomains as shown in ***Figure2*** and further specialize these domains to finer levels as shown in ***Figure3***.

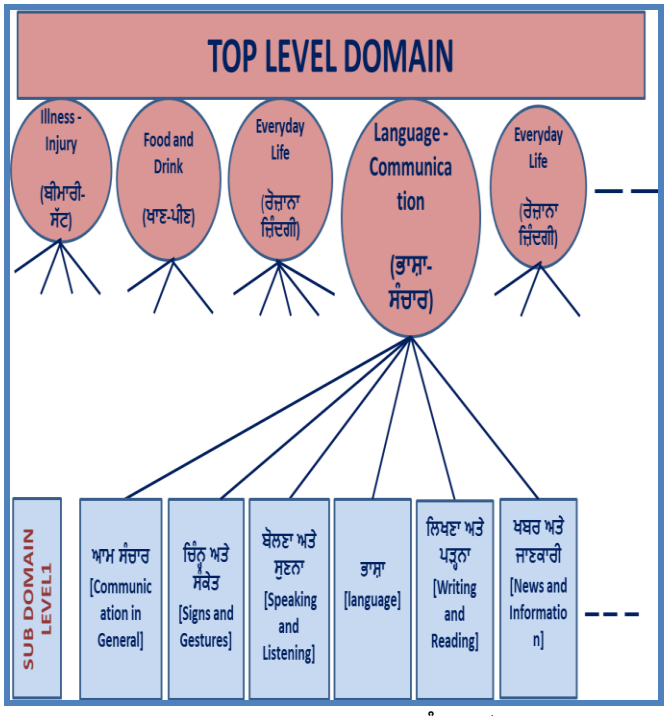

Figure 2: Sub-Domains for "ਭਾਸ਼ਾ-ਸੰਚਾਰ *(Language-Communication)*" Top-Level Domain

The ***Figure2*** here represents the sub-domains for "*ਭਾਸ਼ਾ-ਸੰਚਾਰ (Language-Communication)"* domain. *For example*, sub domains under this top level domain are subdomain "*ਬੋਲਣਾ ਅਤੇ ਸੁਣਨਾ* **(***Speaking and Listening)"* and sub domain *"ਖਬਰ ਅਤੇ ਜਾਣਕਾਰੀ (News and Information)"*. It is worth mentioning that we have shown only some of these subdomains in the figure as the actual list can be much longer encompassing many domains as such.

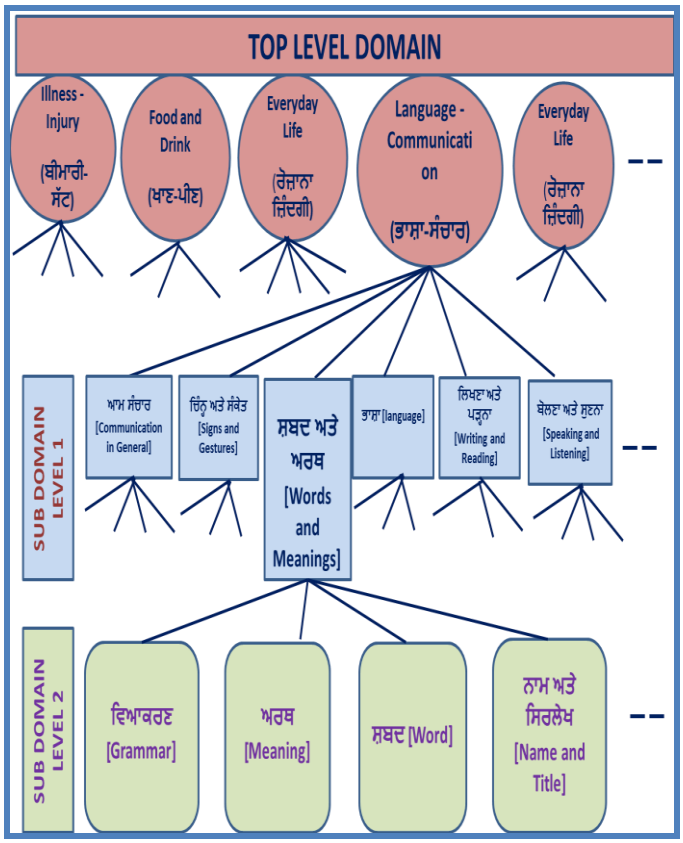

Figure 3: Sub-Domains for "ਸ਼ਬਦ ਅਤੇ ਅਰਥ [Words and Meanings]" Domain under "ਭਾਸ਼ਾ-ਸੰਚਾਰ *(Language-Communication)"* Top-Level Domain

Similarly in ***Figure 3,*** we show some sub-domains for "*ਸ਼ਬਦ ਅਤੇ ਅਰਥ (Words and Meanings),* which is itself a sub domain under top level domain "*ਭਾਸ਼ਾ-ਸੰਚਾਰ (Language-Communication)".* In these figures, the top level domains are represented by oval shape and sub-domains are represented by rectangles.

Table 1 shows a three level hierarchy of domains for a specific top level domain "*ਲੜਾਈ ਅਤੇ ਯੁੱਧ (Fighting and War)*". We clearly observe the controlled hierarchy and fine granularity of domains in the table for this particular domain. *For example*, "*ਥਲ ਸੇਨਾ (Army)*", is a third level domain under its parent domain "*ਹਥਿਆਰਬੰਦ ਫੌਜ (Armed Forces)*" which is a sub-domain of top level domain "*ਲੜਾਈ ਅਤੇ ਯੁੱਧ (Fighting and War)*"

Table1: A hierarchy of domains for a specific Top Level Domain "**ਲੜਾਈ ਅਤੇ ਯੁੱਧ (Fighting and War)**"

ਲੜਾਈ ਅਤੇ ਯੁੱਧ									Top Level Domain
ਆਮ ਲੜਾਈ	ਮਾਰਨਾ ਕੱਟਣਾ	ਜੰਗ	ਹਥਿਆਰਬੰਦ ਫੌਜ	ਆਮ ਹਥਿਆਰ	ਘਾਤਕ ਹਥਿਆਰ ਅਤੇ ਗੋਲਾ ਬਾਰੂਦ	ਦੁਸ਼ਮਣ	ਰੱਖਿਆ	ਹੱਤਿਆ	Sub Domain (Level 1)
•ਲੜਾਈ •ਲੜਾਈ ਦੀ ਸੁਰੂਆਤ ਅਤੇ ਹਮਲਾ •ਮਾਰਨਾ ਅਤੇ ਜਖਮੀ ਕਰਨਾ •ਬਚਾਓ •ਲੜਾਈ ਦਾ ਅੰਤ, ਜਿੱਤ ਹਾਰ	•ਜਾਣ ਬੁਝ ਕੇ ਮਾਰਨਾ •ਅਨਜਾਣੇ ਵਿਚ ਜਾਂ ਦੁਰਘਟਨਾਵਸ ਮਾਰਨਾ	•ਯੁੱਧ •ਫੌਜ •ਜੰਗ ਦਾ ਆਰੰਭ •ਜੰਗ ਲੜਨਾ •ਜਿੱਤ ਹਾਰ •ਜੰਗ ਦਾ ਖਾਤਮਾ ਅਤੇ ਅਮਨ •ਯੁੱਧ-ਹੋਰ	•ਫੌਜ ਦੇ ਲੋਕ •ਫੌਜ ਦਾ ਸੰਗਠਨ •ਸਿਪਾਹੀਆਂ ਦੀਆਂ ਵਸਤੂਆਂ •ਫੌਜ ਸੰਦਾ ਸੰਬੰਧੀ •ਥਲ ਸੈਨਾ •ਹਵਾਈ ਸੈਨਾ •ਸਮੰਦਰੀ ਸੈਨਾ •ਸੈਨਾ ਭਾਈਵਾਲ ਸੇਵਾਂਦਾ	•ਹਥਿਆਰ •ਹਥਿਆਰਾਂ ਸੰਬੰਧੀ	•ਹਥਿਆਰਾਂ ਦੀ ਕਿਸਮਾਂ •ਹਥਿਆਰਾਂ ਦੀ ਵਰਤੋਂ •ਹਥਿਆਰਾਂ ਦੇ ਹਿੱਸੇ •ਗੋਲਾ ਬਾਰੂਦ ਦੀ ਕਿਸਮਾਂ •ਗੋਲਾ ਬਾਰੂਦ ਦੀ ਵਰਤੋਂ	•ਅੰਦਰੂਨੀ ਦੁਸ਼ਮਣ •ਬਾਹਰੀ ਦੁਸ਼ਮਣ	•ਸਰਹੱਦ ਦੀ ਰੱਖਿਆ •ਜਾਨਮਾਲ ਦੀ ਰੱਖਿਆ •ਅੰਦਰੂਨੀ ਰੱਖਿਆ •ਬਾਹਰੀ ਰੱਖਿਆ •ਰੱਖਿਆ- ਹੋਰ	•ਲੋਕਾਂ ਨੂੰ ਮਾਰਨਾ •ਮਾਰਨ ਦੇ ਤਰੀਕੇ	Sub Domain (Level 2)

This organization of data is represented in terms of PUDO ontology as explained below:

C. PUDO Schema

Domains knowledge organization related to Punjabi language is conceptualized by introduction of several key concepts (classes) and relationships (object properties) in PUDO. These concepts in terms of ontology classes, organized and represented in a hierarchical manner, and the linkages amongst these classes in terms of object properties form the ontology schema (PUDO Schema).

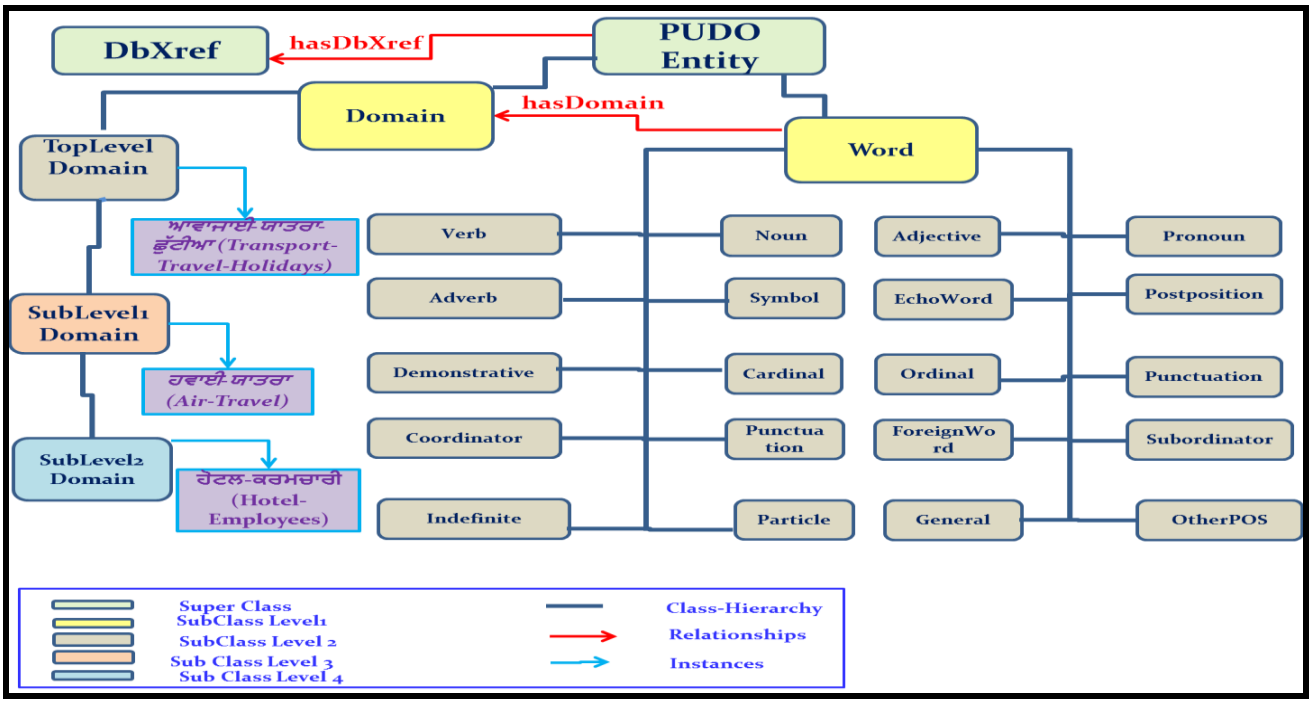

Figure x: A section of the Punjabi Domains Ontology (PUDO) schema showing some key concepts and relationships.

The schema is actually created having knowledge conceived, organized and represented in a manner analogous to a domain expert. The schema is outputted in the form of Web Ontology Language (OWL) language.

Some of the major classes and sub-classes included in the design of PUDO are shown in the ***Figure 4***. Figure 4 also shows the relationships among the PUDO classes providing further important knowledge about the related classes. Furthermore, each PUDO class is described by a set of specific characteristics. For example, the important characteristics of the Domain class include its name, definition, source etc.

Thereafter, the population of the data of Punjabi words takes place which will be assigned domains from the "Domains" class.

D. PUDO Data Population

The PUDO has been populated with data, especially the Punjabi vocabulary, acquired from multiple data sources and having extensive manual work of our team. The acquired data has been represented as instances in the Ontology as described in ***Figure 5***.

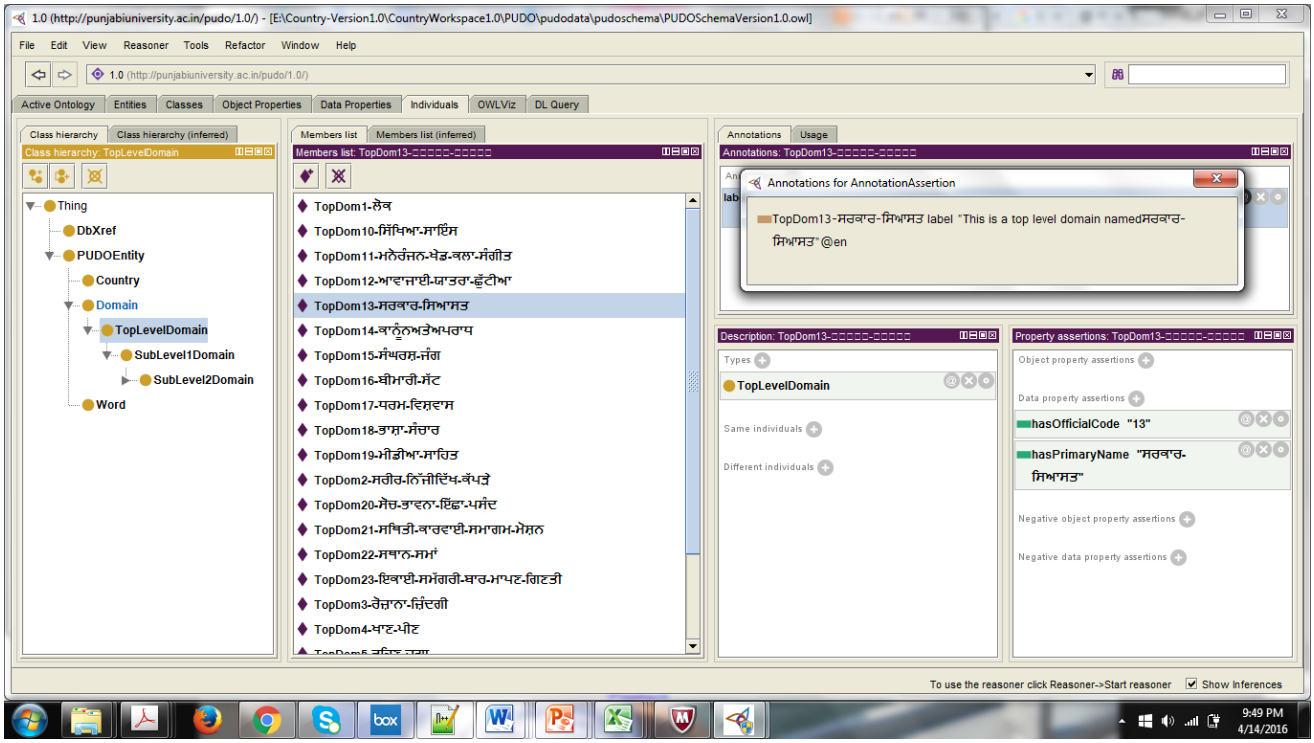

Figure 5: A snapshot of PUDO Ontology Fragment Showing Top Level Domains in Protégé Editor

.

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
xmlns:pudo="http://punjabiuniversity.ac.in/pudo/1.0
/#"    xmlns:rdfs="http://www.w3.org/2000/01/rdf-
schema#" >
  <rdf:Description
rdf:about="http://punjabiuniversity.ac.in/pudo/1.0/
#TopDom1-ਲੋਕ">
<pudo:hasPrimaryName>ਲੋਕ</pudo:hasPrimaryName>
<pudo:hasOfficialCode>1</pudo:hasOfficialCode>
    <rdfs:label xml:lang="en">This is a top level
domain named ਲੋਕ</rdfs:label>
    <rdf:type
rdf:resource="http://punjabiuniversity.ac.in/pudo/1
.0/#TopLevelDomain"/>
  </rdf:Description>
```

Figure 6: An Ontology Fragment in OWL (An Excerpt from PUDO OWL file).

A specialized software system using the Java programming language has been developed to automatically populate PUDO from the data sources. Jena, a widely used Java-based API for parsing, creating and querying OWL ontologies has been integrated in our software system. The functions performed by the customized software include all of the required functions for creating the ontology with its schema and for automatic population. The functions are performed by software for data acquisition, parsing and processing, the formation of instances and linkages between them using the properties defined in the PUDO schema

The populated ontology is generated in OWL, the World Wide Web Consortium's recommended language for ontology authoring and sharing. A small excerpt from the OWL ontology file for PUDO is shown in **Figure 6**.

Eventually we have a corpus having finer level granular domains assigned to words with specificity to Punjabi

ISSN: 2229-6913 (Print), ISSN: 2320-0332 (Online)

language. The ***Table 2*** shows an example case of assignment of words to hierarchical or taxonomical levels of domains under a specific domain "***ਲੜਾਈ ਅਤੇ ਯੁੱਧ (Fighting and War)***".

Table 2: Showing Assignment of Individual Words to Domains for Top Level Domain "**ਲੜਾਈ ਅਤੇ ਯੁੱਧ (Fighting and War)**"

Sr.No.	Top Level Domain	Sub Domain (Level 1)	Sub Domain (Level 2)	Assigned Words Samples
1.	ਲੜਾਈ ਅਤੇ ਯੁੱਧ	ਜੰਗ	ਜੰਗ ਲੜਨਾ	ਹਮਲਾ ਕਰਨਾ, ਘੇਰਾ ਪਾਉਣਾ, ਫੜਨਾ, ਧਾਵਾ - (ਕਿਰਿਆ) ਹਮਲਾਵਰ, ਹਵਾਈ ਹਮਲੇ, ਜੰਗ ਦਾ ਮੈਦਾਨ, ਯੁੱਧ ਰਣਨੀਤੀ -(ਨਾਓ) ਰਣਨੀਤਕ -(ਵਿਸ਼ੇਸ਼ਣ) ਆਦਿ
2.	ਲੜਾਈ ਅਤੇ ਯੁੱਧ	ਜੰਗ	ਜਿੱਤ ਹਾਰ	ਜਿੱਤ, ਹਾਰ, ਸਮਰਪਣ, ਜੇਤੂ -(ਨਾਓ), ਝੁਕਾਉਣਾ, ਯੁੱਧ ਜਿੱਤਣਾ, ਯੁੱਧ ਹਾਰਨਾ, ਆਤਮ ਸਮਰਪਣ ਕਰਨਾ, ਆਤਮ ਸਮਰਪਣ ਕਰਾਉਣਾ, ਮੈਦਾਨ ਛੱਡ ਕੇ ਭੱਜਣਾ, ਜਿਤਾਉਣਾ, ਹਰਾਉਣਾ -(ਕਿਰਿਆ) ਆਦਿ
3.	ਲੜਾਈ ਅਤੇ ਯੁੱਧ	ਘਾਤਕ ਹਥਿਆਰ ਅਤੇ ਗੋਲਾ ਬਾਰੂਦ	ਗੋਲਾ ਬਾਰੂਦ ਦੀ ਕਿਸਮਾਂ	ਬੰਬ, ਗੋਲਾ, ਬਾਰੂਦ, ਗੋਲਾ ਬਾਰੂਦ, ਸੁਰੰਗੀ ਬੰਬ, ਖਾਨ ਬੰਬ, ਸਮਾਂ ਬੰਬ, ਆਰ. ਡੀ. ਐਕਸ, ਵਿਸਫੋਟਕ ਬੰਬ, ਵਿਸਫੋਟਕ ਪਦਾਰਥ, ਵਿਸਫੋਟਕ ਸਮਗਰੀ, ਕਾਰ ਬੰਬ ਆਦਿ -(ਨਾਓ) ਵਿਸਫੋਟਕ, ਬਾਰੂਦੀ, ਮਾਰੂ, ਜਾਨਲੇਵਾ-(ਵਿਸ਼ੇਸ਼ਣ)
4.	ਲੜਾਈ ਅਤੇ ਯੁੱਧ	ਹਥਿਆਰਬੰਦ ਫੌਜ	ਫੌਜ ਦਾ ਸੰਗਠਨ	ਸੈਨਾ ਭਵਨ, ਬਟਾਲੀਅਨ, ਪਲਟਨ, ਰੇਜ਼ਮੈਂਟ, ਬ੍ਰਿਗੇਡ, ਫੌਜੀ ਦਸਤਾ, ਫੌਜੀ ਜਥਾ, ਛਾਉਣੀ, ਕੰਪਨੀ, ਲਸ਼ਕਰ, ਲਾਮ-ਲਸ਼ਕਰ, ਹੈਡਕੁਆਟਰ, ਅਧਾਰ ਕੈਂਪ, ਮੈਸ, ਫੌਜੀ ਦਲ-(ਨਾਓ) ਆਦਿ

IV. EVALUATION

Ontology development requires evaluation of resulting ontology so that we can determine that whether it does serve its purpose correctly 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.

The two approaches have been planned to evaluate the accuracy of PUDO content: (i) a manual approach in which we cross check the randomly selected set of instances and relationships among them from PUDO with content from original sources (ii) a query-based approach in which PUDO data can be queried for information by using a ontology query language SPARQL [16] and that information can easily be cross validated with data from original sources.

V. CONCLUSION AND FUTURE WORK

The present work of ours is a novel effort to conceptualize domains for Punjabi language to finer levels of granularity and represent them in terms of controlled vocabulary in form of an ontology. This is a step ahead from other domains classification works in the area which are more focused on specialization of some concepts around words rather than having generalized concepts and that too without hierarchy.

The compendium of knowledge represented in PUDO can be used for a variety of applications such as summarization, topic classification, data mining, text mining, semantic annotation and other NLP applications involving semantics. The queries, formulated in ontology query languages such as SPARQL, can provide a platform for these tasks. Fine granularity of domains available in PUDO presents an opportunity and support for more sophisticated semantics.

Our focus presently and in future is to extend the coverage of words to make it a comprehensive and community accepted resource. The effort is also on for providing a platform to leverage the knowledge contained in PUDO by enabling the browsing and navigation accessible from '***http://punjabiuniversity.ac.in/pbiuniweb/pages/departments/newcomputerscience/semtadagroup/pudo.html***'. Thereafter, the effort would be to build some basic applications focusing the knowledge of PUDO to test its utility for semantic applications.

References

[1] Liddy, E. D. (2001). Natural language processing. In: Encyclopedia of Library and Information Science, 2nd edn. Marcel Decker, Inc., New York.

[2] Nirenburg, S., McShane, M., & Beale, S. (2004). The rationale for building resources expressly for NLP. In *Proc. of the 4th International Conference on Language Resources and Evaluation* (Vol. 218).

[3] Gosal, G. P. S., Kannan, N., & Kochut, K. J. (2011, November). ProKinO: a framework for protein kinase ontology. In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on* (pp. 550-555). IEEE.

[4] Punjabi Language, [29 March 2016, Date last accessed]; Available from: https://en.wikipedia.org/wiki/Punjabi_language and http://www.britannica.com/topic/Punjabi-language.

[5] Let us Learn Punjabi. [29 March 2016, Date last accessed]; Available from: http://www.learnpunjabi.org/intro1.asp.

[6] Gardner, S. P. (2005). Ontologies and semantic data integration. *Drug discovery today*, *10*(14), 1001-1007.

[7] 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.

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

[9] Spyns, P., Meersman, R., & Jarrar, M. (2002). Data modelling versus ontology engineering. *ACM SIGMod Record*, *31*(4), 12-17.

[10] Motik, B., Horrocks, I., & Sattler, U. (2009). Bridging the gap between OWL and relational databases. *Web Semantics: Science, Services and Agents on the World Wide Web*, *7*(2), 74-89.

[11] Gosal, G., Ontology Building: An Integrative View of Methodologies, International Journal on Recent and Innovation Trends in Computing and Communication, July- 2015, pp. ISSN: 2321-8169 Volume: 3 Issue: 7

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

[13] Punjabi Topic Dictionary, [31 March 2016, Date last accessed]; Available from: http://www.learnpunjabi.org/vocabulary/vocabulary1.asp?id=46

[14] Jena Ontology API, , [31 March 2016, Date last accessed]; Available from: https://jena.apache.org/documentation/ontology/

[15] Web Ontology Language (OWL), [31 March 2016, Date last accessed]; Available from: http://www.w3.org/TR/owl-ref/

[16] SPARQL, [31 March 2016, Date last accessed]; Available from: (http://www.w3.org/TR/rdf-sparql-query/)

ISSN: 2229-6913 (Print), ISSN: 2320-0332 (Online)