IndoWordNet Conversion to Web Ontology Language (OWL)

Apurva Nagvenkar
DCST, Goa University
apurv.nagvenkar@gmail.com

Jyoti Pawar
DCST, Goa University
jyotidpawar@gmail.com

Pushpak Bhattacharyya
CSE, IIT Bombay
pb@cse.iitb.ac.in

Abstract

WordNet plays a significant role in Linked Open Data (LOD) cloud. It has numerous application ranging from ontology annotation to ontology mapping. IndoWordNet is a linked WordNet connecting 18 Indian language WordNets with Hindi as a source WordNet. The Hindi WordNet was initially developed by linking it to English WordNet.

In this paper, we present a data representation of IndoWordNet in Web Ontology Language (OWL). The schema of Princeton WordNet has been enhanced to support the representation of IndoWordNet. This IndoWordNet representation in OWL format is now available to link other web resources. This representation is implemented for eight Indian languages.

1 Introduction

The World Wide Web (WWW) has formed a revolution in the data availability there is no other place in the world where we can find so much of the information, but the current web structure fails to make best out of it. The user can access limitless data from the web yet, it becomes a tedious task to retrieve relevant information. Data available on the Web covers diverse structures, formats and content. It also lacks a uniform organization of scheme that would allow easy access of data and information (Candan et al., 2001). Many frameworks have been proposed to support the search engine and information access. Resource Description Framework[1](RDF), Web Ontology Language[2](OWL) is one of the framework which provides a platform for standardization and organization of data from the Web. It has been highly influenced by the web standards community.

	Noun	Verb	Adjective	Adverb	Total
Bengali	27281	2804	5815	445	36346
Gujarati	26503	2805	5828	445	35599
Hindi	29106	3306	6178	482	39072
Kashmiri	21041	2660	5365	400	29469
Konkani	23144	3000	5744	482	32370
Odia	27216	2418	5273	377	35284
Punjabi	23255	2836	5830	443	32364
Urdu	22990	2801	5786	443	34280

Table 1: POS wise statistics for Indradhanush

WordNet (Fellbaum, 1998), a lexical knowledge base system that has been adopted by the Semantic Web research community. The current essential need is to link WordNet with different resources in order to assist Natural Language Processing applications. IndoWordNet (Bhattacharyya, 2010) is an Indian community which builds WordNets for Indian languages. It is a multilingual WordNet which links WordNets of different Indian languages on a common identification number called as synset_id given to each concept (Bhattacharyya, 2010). It is constructed using the expansion model where Hindi WordNet synsets are taken as a source. The concepts provided along with the Hindi synsets are first conceived and appropriate concepts in target language are manually provided by the language experts. Figure 1 shows the statistics of Indradhanush Consortium which consist seven Indian languages belonging to Indo-Aryan family and is part of IndowordNet Consortium.

To use WordNet in Semantic Web the data model for WordNet should be extensible, interoperable and flexible. It was created as a semantic network of word meanings which at the conceptual level is a directed graph with labeled nodes and arcs (Graves and Gutierrez, 2006). Hence, OWL can be used to model WordNet since, it facilitates data manipulations and queries over the

[1] http://www.w3.org/RDF
[2] http://www.w3.org/TR/owl-features

graph structure. The main objective of this paper is to represent IndoWordNet to OWL representation.

The rest of the paper is organized as follows section 2 describes the related work. Section 3 introduces to Semantic Web Layer Cake Model. Section 4 presents the architecture of IndoWordNet OWL; section 5 gives the implementation details, followed by conclusion and future work.

2 Related Work

WordNets other than Indian languages are already available in RDF form. The work on Princeton WordNet (Assem et al., 2006) conversion to RDF/OWL was carried out by WordNet Task Force[3]. The main goal of this conversion was to represent a language in use of Semantic Web community and to provide application developers a resource. Also, the representation was done in such a way that it maintained the WordNets conceptual model.

There are other projects focusing on lexical meta-models. Lexical Markup Framework (LMF) (Francopoulo et al., 2009). IndoWordNet is already available in this format by IndoNet (Bhatt et al., 2013) which proposes modification to LMF to integrate Universal Word Dictionary (Uchida et al., 1999) and Suggested Upper Merged Ontology (SUMO) (Pease et al., 2002).

3 Semantic Web Layer Cake Model

The Semantic Web is not a separate web but a vision for the future of the Web where information is given explicit meaning which makes easier for machine to automatically process and integrate the information available on the web. OWL is a part of the growing stack of W3C recommendations related to the semantic web (McGuinness and Harmelen, 2004).

Figure 1. is the semantic web layer cake model (Hendler, 2001). This model is divided into three section:

1. Hypertext Web technologies: The bottom layer contains technologies which are used by hypertext web that includes Unicode, Universal Resource Indicator (URI), XML and XML-schema. Unicode is used to represent and manipulate text for different languages. URI represents the resources uniquely. XML provides the syntax for structured document, but does not provide any meaning to the document. XML schema restricts the structure of the document and extends XML with datatypes.

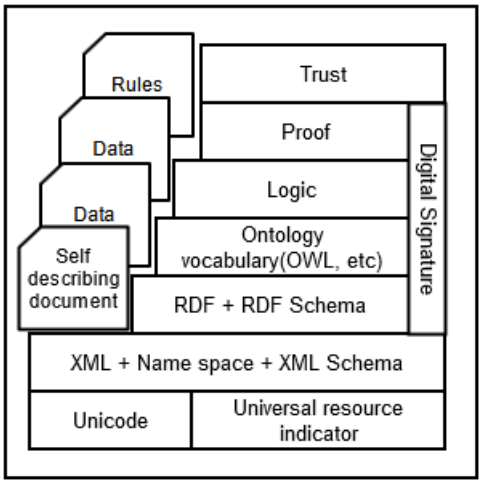

Figure 1: Semantic web layer cake model

2. Standardized Semantic Web technologies: The middle layer contains technologies which are already standardized by Semantic Web community that includes RDF, RDFS, OWL and SPARQL. RDF is a data model to represent triple, i.e. objects and relationship between them. It provides simple semantics and is represented by XML syntax. RDF schema can be viewed as an extensible, object oriented type system based on RDF (Huang and Zhou, 2007). OWL is an envelope to the RDF schema and enriches the expressibility of the RDF schema by expressing more properties like transitivity, symmetry, cardinality, etc.

3. Unrealized Semantic Web technologies: The top layer contains technologies like digital signatures, trust, proof, etc this technologies are not yet standardized by Semantic Web community and needs to be implemented in order to realize Semantic Web.

4 OWL for IndoWordNet

The architecture of the IndoWordNet OWL representation is adopted from WordNet Task Force (Assem et al., 2006). The architecture of IndoWordNet OWL contains three main classes i.e.

[3] http://www.w3.org/TR/wordnet-rdf/

Synset[4], WordSense and Word[5] .

The schema for representing IndoWordNet[6] using OWL is shown in figure 2 below.

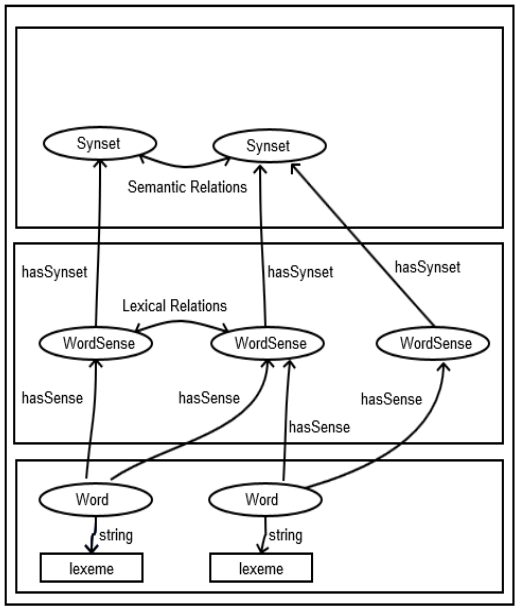

Figure 2: IndoWordNet OWL schema

The schema includes three layers, namely Concept layer, WordSense layer and Word layer which are previously described in (Huang and Zhou, 2007). Every synset has a unique concept and can have several words associated with it sharing the same concept. WordSense represents a unique sense of a word. It is also possible to represent a word with many WordSenses. IndoWordNet OWL schema handles the relations by dividing them into properties, i.e. Semantic property and Lexical property. Semantic property represents the semantic relations which are handled in concept layer, whereas lexical property represents lexical relations which are handled in WordSense layer. All the remaining types of semantic relations and lexical relations become the sub property of semantic and lexcial property. The above schema uses several predicates[7] i.e. properties.

IndoWordNet OWL schema elaborates the semantic relationship like meronymy and holonymy by classifying them into the sub properties based on their attributes[8] whereas in Princeton WordNet there is no such division.

In IndoWordNet OWL, the RDF files are organized in such a way that the management is done systematically. Unlike (Assem et al., 2006) all the RDF files are placed in one directory.

Following is the formatting of URIs for IndoWordNet:

- URI representation of a synset: `http://nlp.unigoa.ac.in/indonet/owl/hindi/v1/synset/noun/24.rdf`
- URI representation of a wordSense: `http://nlp.unigoa.ac.in/indonet/owl/hindi/v1/wordSense/ł/noun/1930.rdf`
- URI representation of a word: `http://nlp.unigoa.ac.in/indonet/owl/hindi/v1/word/ł.rdf`

5 Implementation Details

The IndoWordNet OWL is currently available in seven Indian languages. It is developed using JAVA platform, using Apache Jena[9] and IndoWordNet Application Programming Interface(API). The above architecture can be used by other Indian langauges to represent their respective wordNets in OWL format. The repository of IndoWordNet OWL is available on `http://nlp.unigoa.ac.in/indonet/owl/`.

6 Conclusion and Future Work

The heart of Semantic Web is Linked Data that provides integration and reasoning of the data on web. The representation of IndoWordNet to OWL will facilitate the semantic web community as the WordNet is strong lexical resource that has strengthened, enlarged and build up the other resources because of its taxonomy. In this paper we have presented the framework to represent the Indian wordNets in the OWL format. Currently, we have represented eight Indian language WordNets in OWL format. In future, we will like to represent the WordNets from other Indian languages in OWL format. Following are some future work to this problem.

[4] `http://nlp.unigoa.ac.in/indonet/owl/web/syn.php`

[5] `http://nlp.unigoa.ac.in/indonet/owl/web/wdSenseAndWord.php`

[6] `http://nlp.unigoa.ac.in/indonet/owl/IndoWNetSchema.rdf`

[7] `http://nlp.unigoa.ac.in/indonet/owl/web/prop.php`

[8] `http://nlp.unigoa.ac.in/indonet/owl/web/propdist.php`

[9] `https://jena.apache.org/`

Interlinking of WordNets: As the IndoWordNet is developed using ILI. The advantage of this approach is that it preserves the semantic structure, but it also has some disadvantages. The drawbacks of this approach are lexical gap and semantic gap (Fellbaum and Vossen, 2012). As a result, an effort must be made to interlink the WordNet using Common Concept Hierarchy (Bhatt et al., 2013) as a backbone to link lexicons of different languages.

Need of approach to link DBpedia: The work on linking the IndoWordNet to DBpedia should be carried out as, DBpedia is the nucleus for the web of data and most of the resources are already linked to DBpedia.

Link it to other Resources: We expect that use of the OWL representation of IndoWordNet will be used as an infrastructure to enrich and link other web resources in India.

References

[Graves and Gutierrez2006] Alvaro Graves, Claudio Gutierrez. 2006. *Data Representation for WordNet: A Case for RDF*. 3rd Global WordNet Association Conference.

[Bhatt et al.2013] Brijesh Bhatt, Lahari Poddar, Pushpak Bhattacharyya. 2013. *IndoNet: A Multilingual Lexical Knowledge Network for Indian Languages*. Association for Computational Linguistics.

[Fellbaum1998] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database.*. Cambridge, MA: MIT Press.

[Fellbaum and Vossen2012] Christiane Fellbaum and Piek Vossen. 2012. *Challenges for a multilingual wordnet.*. Lang. Resour. Eval., 46(2):313326.

[McGuinness and Harmelen2004] Deborah L. McGuinness, Frank van Harmelen. 2004. *OWL Web Ontology Language*. http://www.w3.org/TR/owl-features.

[Francopoulo et al.2009] Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2009. *LMF for Multilingual Specialized Lexicons*. LREC Workshop on Acquiring and Representing Multilingual, Specialized Lexicons.

[Hendler2001] J. Hendler. 2001. *Agents and the Semantic Web*. IEEE Intellegent Systems.

[Candan et al.2001] K. Seluk Candan, Huan Liu, and Reshma Suvarna. 2001. *Resource description framework: metadata and its applications*. SIGKDD Explor. Newsl. 3, 1 (July 2001), 6-19.

[Kuroda et al.2010] Kow Kuroda, Francis Bond, Kentaro Torisawa. 2010. *Why Wikipedia needs to make friends with WordNet*. 5th Global WordNet Association Conference.

[Assem et al.2006] Mark van Assem, Aldo Gangemi, Guus Schreiber. 2006. *Conversion of WordNet to a standard RDF/OWL representaion*. Proceedings of LERC.

[Casado et al.2005] Maria Ruiz-Casado, Enrique Alfoneseca, Pablo Castells. 2005. *Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets*. In: Proceedings of the Atlantic Web Intelligence Conference, AWIC-2005.

[Bhattacharyya2010] Pushpak Bhattacharyya. 2010. *IndoWordNet*. Proceedings of LERC.

[Auer et al.2007] Soren Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, Zachary Ives. 2007. *DBpedia: A Nucleus for a Web of Open Data*. ISWC'07/ASWC'07 Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference.

[Huang and Zhou2007] Xiao-xi Huang, Chang-le Zhou. 2007. *An OWL-based WordNet lexical ontology*. Journal of Zhejiang Science A.

[Pease et al.2002] Adam Pease, Ian Niles and John Li 2002. *The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications*. In Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web.

[Uchida et al.1999] H. Uchida, M. Zhu, and T. Della Senta 1999. *UNL- a Gift for the Millenium*. United Nations University Press, Tokyo.