IndoNet: A Multilingual Lexical Knowledge Network for Indian Languages

Brijesh Bhatt Lahari Poddar Pushpak Bhattacharyya
Center for Indian Language Technology
Indian Institute of Technology Bombay
Mumbai, India
{ brijesh, lahari, pb } @cse.iitb.ac.in

Abstract

We present IndoNet, a multilingual lexical knowledge base for Indian languages. It is a linked structure of wordnets of 18 different Indian languages, Universal Word dictionary and the Suggested Upper Merged Ontology (SUMO). We discuss various benefits of the network and challenges involved in the development. The system is encoded in Lexical Markup Framework (LMF) and we propose modifications in LMF to accommodate Universal Word Dictionary and SUMO. This standardized version of lexical knowledge base of Indian Languages can now easily be linked to similar global resources.

1 Introduction

Lexical resources play an important role in natural language processing tasks. Past couple of decades have shown an immense growth in the development of lexical resources such as wordnet, Wikipedia, ontologies etc. These resources vary significantly in structure and representation formalism.

In order to develop applications that can make use of different resources, it is essential to link these heterogeneous resources and develop a common representation framework. However, the differences in encoding of knowledge and multilinguality are the major road blocks in development of such a framework. Particularly, in a multilingual country like India, information is available in many different languages. In order to exchange information across cultures and languages, it is essential to create an architecture to share various lexical resources across languages.

In this paper we present IndoNet, a lexical resource created by merging wordnets of 18 different Indian languages[1], Universal Word Dictionary (Uchida et al., 1999) and an upper ontology, SUMO (Niles and Pease, 2001).

Universal Word (UW), defined by a headword and a set of restrictions which give an unambiguous representation of the concept, forms the vocabulary of Universal Networking Language. Suggested Upper Merged Ontology (SUMO) is the largest freely available ontology which is linked to the entire English WordNet (Niles and Pease, 2003). Though UNL is a graph based representation and SUMO is a formal ontology, both provide language independent conceptualization. This makes them suitable candidates for interlingua.

IndoNet is encoded in Lexical Markup Framework (LMF), an ISO standard (ISO-24613) for encoding lexical resources (Francopoulo et al., 2009).

The contribution of this work is twofold,

1. We propose an architecture to link lexical resources of Indian languages.

2. We propose modifications in Lexical Markup Framework to create a linked structure of multilingual lexical resources and ontology.

2 Related Work

Over the years wordnet has emerged as the most widely used lexical resource. Though most of the wordnets are built by following the standards laid by English Wordnet (Fellbaum, 1998), their conceptualizations differ because of the differences in lexicalization of concepts across languages. 'Not

[1] Wordnets for Indian languages are developed in IndoWordNet project. Wordnets are available in following Indian languages: Assamese, Bodo, Bengali, English, Gujarati, Hindi, Kashmiri, Konkani, Kannada, Malayalam, Manipuri, Marathi, Nepali, Punjabi, Sanskrit, Tamil, Telugu and Urdu. These languages covers 3 different language families, Indo Aryan, Sino-Tebetian and Dravidian.http://www.cfilt.iitb.ac.in/indowordnet

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 268–272,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

only that, there exist lexical gaps where a word in one language has no correspondence in another language, but there are differences in the ways languages structure their words and concepts'. (Pease and Fellbaum, 2010).

The challenge of constructing a unified multilingual resource was first addressed in EuroWordNet (Vossen, 1998). EuroWordNet linked wordnets of 8 different European languages through a common interlingual index (ILI). ILI consists of English synsets and serves as a pivot to link other wordnets. While ILI allows each language wordnet to preserve its semantic structure, it has two basic drawbacks as described in Fellbaum and Vossen (2012),

1. An ILI tied to one specific language clearly reflects only the inventory of the language it is based on, and gaps show up when lexicons of different languages are mapped to it.

2. The semantic space covered by a word in one language often overlaps only partially with a similar word in another language, resulting in less than perfect mappings.

Subsequently in KYOTO project[2], ontologies are preferred over ILI for linking of concepts of different languages. Ontologies provide language indpendent conceptualization, hence the linking remains unbiased to a particular language. Top level ontology SUMO is used to link common base concepts across languages. Because of the small size of the top level ontology, only a few wordnet synsets can be linked directly to the ontological concept and most of the synsets get linked through subsumption relation. This leads to a significant amount of information loss.

KYOTO project used Lexical Markup Framework (LMF) (Francopoulo et al., 2009) as a representation language. 'LMF provides a common model for the creation and use of lexical resources, to manage the exchange of data among these resources, and to enable the merging of a large number of individual electronic resources to form extensive global electronic resources' (Francopoulo et al., 2009). Soria et al. (2009) proposed WordNet-LMF to represent wordnets in LMF format. Henrich and Hinrichs (2010) have further modified Wordnet-LMF to accommodate lexical relations. LMF also provides extensions for multilingual lexicons and for linking external resources, such as ontology. However, LMF does not explicitly define standards to share a common ontology among multilingual lexicons.

Our work falls in line with EuroWordNet and Kyoto except for the following key differences,

- Instead of using ILI, we use a 'common concept hierarchy' as a backbone to link lexicons of different languages.

- In addition to an upper ontology, a concept in common concept hierarchy is also linked to Universal Word Dictionary. Universal Word dictionary provides additional semantic information regarding argument types of verbs, that can be used to provide clues for selectional preference of a verb.

- We refine LMF to link external resources (e.g. ontologies) with multilingual lexicon and to represent Universal Word Dictionary.

3 IndoNet

IndoNet uses a common concept hierarchy to link various heterogeneous lexical resources. As shown in figure 1, concepts of different wordnets, Universal Word Dictionary and Upper Ontology are merged to form the common concept hierarchy. Figure 1 shows how concepts of English WordNet (EWN), Hindi Wordnet (HWN), upper ontology (SUMO) and Universal Word Dictionary (UWD) are linked through common concept hierarchy (CCH).

This section provides details of Common Concept Hierarcy and LMF encoding for different resources.

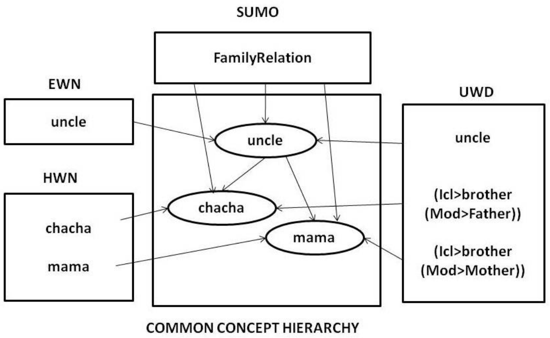

Figure 1: An Example of Indonet Structure

[2] http://kyoto-project.eu/xmlgroup.iit.cnr.it/kyoto/index.html

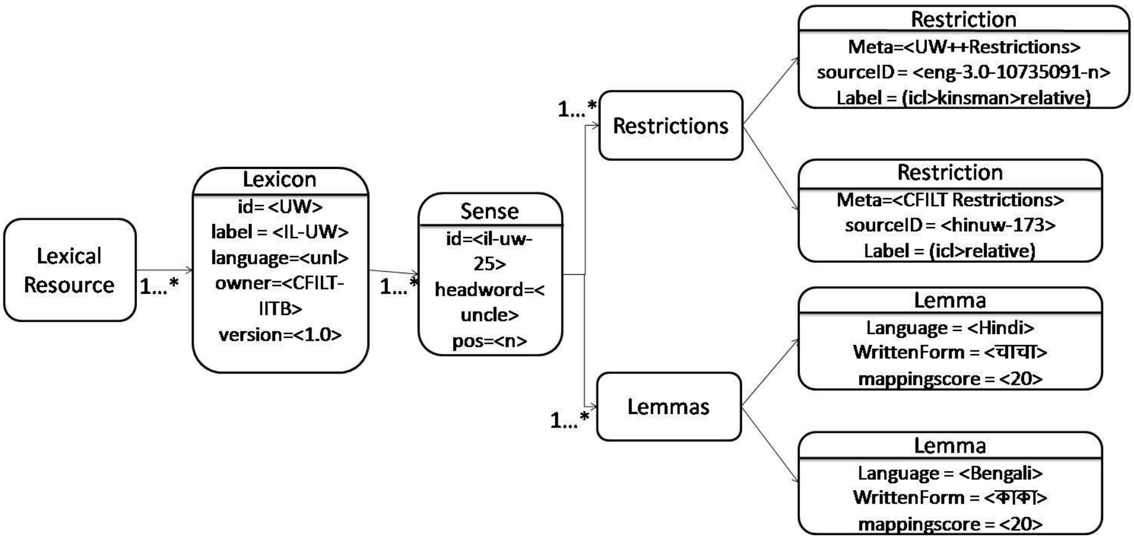

Figure 2: LMF representation for Universal Word Dictionary

3.1 Common Concept Hierarchy (CCH)

The common concept hierarchy is an abstract pivot index to link lexical resources of all languages. An element of a common concept hierarchy is defined as $< sinid_1, sinid_2, ..., uwid, sumoid >$ where, $sinid_i$ is synset id of i^{th} wordnet, uw_id is universal word id, and $sumo\_id$ is SUMO term id of the concept. Unlike ILI, the hypernymy-hyponymy relations from different wordnets are merged to construct the concept hierarchy. Each synset of wordnet is directly linked to a concept in 'common concept hierarchy'.

3.2 LMF for Wordnet

We have adapted the Wordnet-LMF, as specified in Soria et al. (2009). However IndoWordnet encodes more lexical relations compared to EuroWordnet. We enhanced the Wordnet-LMF to accommodate the following relations: *antonym, gradation, hypernymy, meronym, troponymy, entailment* and cross part of speech links for *ability* and *capability*.

3.3 LMF for Universal Word Dictionary

A Universal Word is composed of a headword and a list of restrictions, that provide unique meaning of the UW. In our architecture we allow each sense of a headword to have more than one set of restrictions (defined by different UW dictionaries) and be linked to lemmas of multiple languages with a confidence score. This allows us to merge multiple UW dictionaries and represent it in LMF format. We introduce four new LMF classes; *Restrictions, Restriction, Lemmas and Lemma* and add new attributes; *headword* and *mapping score* to existing LMF classes.

Figure 2 shows an example of LMF representation of UW Dictionary. At present, the dictionary is created by merging two dictionaries, UW++ (Boguslavsky et al., 2007) and CFILT Hin-UW[3]. Lemmas from different languages are mapped to universal words and stored under the *Lemmas* class.

3.4 LMF to link ontology with Common Concept Hierarchy

Figure 3 shows an example LMF representation of CCH. The interlingual pivot is represented through *SenseAxis*. Concepts in different resources are linked to the *SenseAxis* in such a way that concepts linked to same *SenseAxis* convey the same *Sense*.

Using LMF class *MonolingualExternalRefs*, ontology can be integrated with a monolingual lexicon. In order to share an ontology among multilingual resources, we modify the original core package of LMF.

As shown in figure 3, a SUMO term is shared across multiple lexicons via the *SenseAxis*. SUMO is linked with concept hierarchy using the follow-

[3] http://www.cfilt.iitb.ac.in/~hdict/webinterface_user/

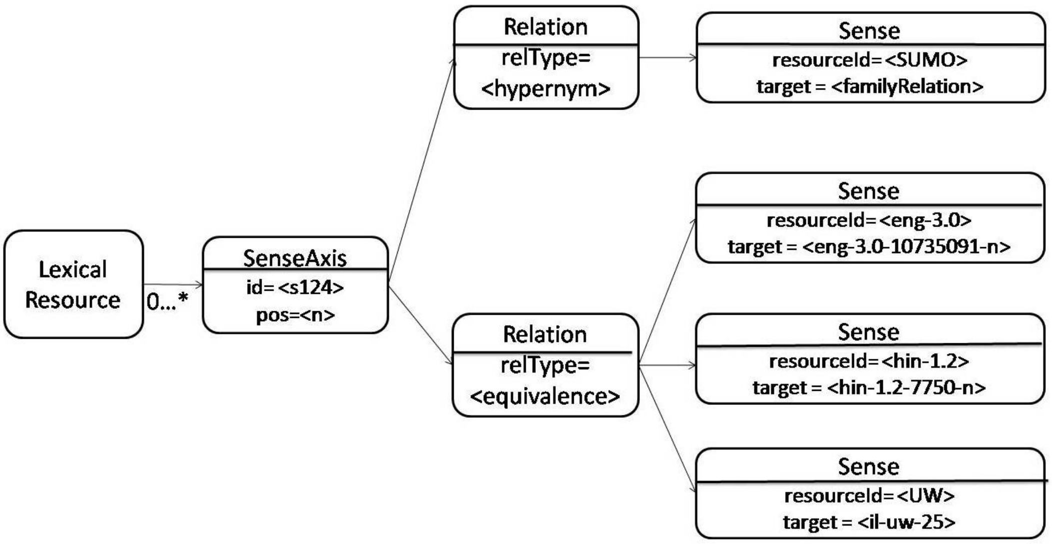

Figure 3: LMF representation for Common Concept Hierarchy

ing relations: *antonym, hypernym, instance and equivalent*. In order to support these relations, *Reltype* attribute is added to the interlingual *Sense* class.

4 Observation

Table 1 shows *part of speech* wise status of linked concepts[4]. The concept hierarchy contains 53848 concepts which are shared among wordnets of Indian languages, SUMO and Universal Word Dictionary. Out of the total 53848 concepts, 21984 are linked to SUMO, 34114 are linked to HWN and 44119 are linked to UW. Among these, 12,254 are common between UW and SUMO and 21984 are common between wordnet and SUMO.

POS	HWN	UW	SUMO	CCH
adjective	5532	2865	3140	5193
adverb	380	2697	249	2813
noun	25721	32831	16889	39620
verb	2481	5726	1706	6222
total	34114	44119	21984	53848

Table 1: Details of the concepts linked

This creates a multilingual semantic lexicon that captures semantic relations between concepts of different languages. Figure 1 demonstrates this with an example of 'kinship relation'. As shown in Figure 1, *'uncle'* is an English language concept defined as 'the brother of your father or mother'. Hindi has no concept equivalent to 'uncle' but there are two more specific concepts *'kaka'*, 'brother of father.' and *'mama'*, 'brother of mother.'

The lexical gap is captured when these concepts are linked to CCH. Through CCH, these concepts are linked to SUMO term *'FamilyRelation'* which shows relation between these concepts. Universal Word Dictionary captures exact relation between these concepts by applying restrictions *[chacha] uncle(icl>brother (mod>father))* and *[mama] uncle(icl>brother (mod>mother))*. This makes it possible to link concepts across languages.

5 Conclusion

We have presented a multilingual lexical resource for Indian languages. The proposed architecture handles the 'lexical gap' and 'structural divergence' among languages, by building a common concept hierarchy. In order to encode this resource in LMF, we developed standards to represent UW in LMF.

IndoNet is emerging as the largest multilingual resource covering 18 languages of 3 different language families and it is possible to link or merge other standardized lexical resources with it.

Since Universal Word dictionary is an integral part of the system, it can be used for UNL based

[4] Table 1 shows data for Hindi Wordnet. Statistics for other wordnets can be found at http://www.cfilt.iitb.ac.in/wordnet/webhwn/iwn_stats.php

Machine Translation tasks. Ontological structure of the system can be used for multilingual information retrieval and extraction.

In future, we aim to address ontological issues of the common concept hierarchy and integrate domain ontologies with the system. We are also aiming to develop standards to evaluate such multilingual resources and to validate axiomatic foundation of the same. We plan to make this resource freely available to researchers.

Acknowledgements

We acknowledge the support of the Department of Information Technology (DIT), Ministry of Communication and Information Technology, Government of India and also of Ministry of Human Resource Development. We are also grateful to Study Group for Machine Translation and Automated Processing of Languages and Speech (GETALP) of the Laboratory of Informatics of Grenoble (LIG) for assissting us in building the Universal Word dictionary.

References

I. Boguslavsky, J. Bekios, J. Cardenosa, and C. Gallardo. 2007. Using Wordnet for Building an Interlingual Dictionary. In *Fifth International Conference Information Research and Applications*, (TECH 2007).

Christiane Fellbaum and Piek Vossen. 2012. Challenges for a multilingual wordnet. *Language Resources and Evaluation*, 46(2):313–326, june.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet, and Claudia Soria. 2009. Multilingual resources for NLP in the lexical markup framework (LMF). *Language Resources and Evaluation*.

Verena Henrich and Erhard Hinrichs. 2010. Standardizing wordnets in the ISO standard LMF: Wordnet-LMF for GermaNet. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 456–464, Stroudsburg, PA, USA.

Ian Niles and Adam Pease. 2001. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9, New York NY USA. ACM.

Ian Niles and Adam Pease. 2003. Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings Of The 2003 International Conference On Information And Knowledge Engineering (Ike 03), Las Vegas*, pages 412–416.

Adam Pease and Christiane Fellbaum. 2010. Formal ontology as interlingua: The SUMO and WordNet linking project and global wordnet. In *Ontology and Lexicon, A Natural Language Processing perspective*, pages 25–35. Cambridge University Press.

Claudia Soria, Monica Monachini, and Piek Vossen. 2009. Wordnet-LMF: fleshing out a standardized format for wordnet interoperability. In *Proceedings of the 2009 international workshop on Intercultural collaboration*, IWIC '09, pages 139–146, New York, NY, USA. ACM.

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

Piek Vossen, editor. 1998. *EuroWordNet: a multilingual database with lexical semantic networks.* Kluwer Academic Publishers, Norwell, MA, USA.