Indian Language Wordnets and their Linkages with Princeton WordNet

Diptesh Kanojia[1,2,3], **Kevin Patel**[1], **Pushpak Bhattacharyya**[1]
[1]IIT Bombay, [2]Monash University,
[3]IITB-Monash Research Academy,
{diptesh, kevin.patel, pb}@cse.iitb.ac.in

Abstract

Wordnets are rich lexico-semantic resources. Linked wordnets are extensions of wordnets, which link similar concepts in wordnets of different languages. Such resources are extremely useful in many Natural Language Processing (NLP) applications, primarily those based on knowledge-based approaches. In such approaches, these resources are considered as gold standard/oracle. Thus, it is crucial that these resources hold correct information. Thereby, they are created by human experts. However, human experts in multiple languages are hard to come by. Thus, the community would benefit from sharing of such manually created resources. In this paper, we release mappings of 18 Indian language wordnets linked with Princeton WordNet. We believe that availability of such resources will have a direct impact on the progress in NLP for these languages.

1. Introduction

Wordnets (Fellbaum, 1998) have been useful in different Natural Language Processing applications such as Word Sense Disambiguation (TufiŞ et al., 2004; Sinha et al., 2006), Machine Translation (Knight and Luk, 1994) *etc.*

Linked Wordnets are extensions of wordnets. In addition to language-specific information captured in constituent wordnets, linked wordnets have a notion of an interlingual index, which connects similar concepts in different languages. Such linked wordnets have found their application in machine translation (Hovy, 1998), cross-lingual information retrieval (Gonzalo et al., 1998), *etc.*

Given the extensive application of wordnets in different NLP applications, creation and maintenance of wordnets involve expert involvement. Such involvement is costly both in terms of time and resources. This is further amplified in case of linked wordnets, where experts need to have knowledge of multiple languages.

India is a vast country with massive language diversity. According to a census in 2001, there are 122 major languages [1], out of which, 29 have more than a million native speakers. The IndoWordNet project contains wordnets of 18 of these languages. These wordnets were created using expansion approach with Hindi Wordnet as the pivot.

This paper makes the following contributions:

- We release the latest version of 18 wordnets under the IndoWordNet project as a single bundle[2].
- Using mappings between Princeton WordNet and Hindi wordnet, we create and release mappings between Princeton WordNet and these 18 languages wordnet.

The rest of the paper is organized as follows: Section 2. covers some background and related work needed for further discussions. Section 3. describes the released resources. Section 4. discusses different issues encountered in the creation of these datasets, followed by the conclusion and future work.

[1]http://en.wikipedia.org/wiki/Languages_of_India
[2]http://www.cfilt.iitb.ac.in/ilw

2. Background and Related Work

Princeton WordNet or the English WordNet was the first wordnet and inspired the development of many other wordnets. EuroWordNet (Vossen and others, 1997) is a linked wordnet comprising of wordnets for European languages, *viz*, Dutch, Italian, Spanish, German, French, Czech and Estonian. Each of these wordnets is structured in the same way as the Princeton WordNet for English (Miller et al., 1990) - synsets (sets of synonymous words) and semantic relations between them. Each wordnet separately captures a language-specific information. In addition, the wordnets are linked to an Inter-Lingual-Index, which uses Princeton WordNet as a base. This index enables one to go from concepts in one language to similar concepts in any other language. Such features make this resource helpful in cross-lingual NLP applications.

IndoWordNet (Bhattacharyya, 2010) is a linked wordnet comprising of wordnets for major Indian languages, *viz*, Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu, and Urdu. These wordnets have been created using the expansion approach with Hindi WordNet as a pivot, which is partially linked to English WordNet. We exploit these links to create mappings from English WordNet to wordnets of other languages.

3. Resources

In this section, we describe the resources released with our work. We release two primary resources with our dataset which are described in subsections 3.1. and 3.2. below.

3.1. Indian Language WordNets

The creation of IndoWordNet began in 2000 with Hindi WordNet. Due to the complex nature of Indian language families, and many other reasons such as morphological richness, gender information etc. it was decided that Hindi be used as a pivot for linking all the Indian Languages. Hindi shares many common features and borrowed concepts from ancient Indian languages like Sanskrit and is the

	Noun	Verb	Adjectives	Adverbs	Total
Assamese	9065	1676	3805	412	14958
Bengali	27281	2804	5815	445	36346
Bodo	8788	2296	4287	414	15785
Gujarati	26503	2805	5828	445	35599
Hindi	29807	3687	6336	541	40371
Kannada	12765	3119	5988	170	22042
Kashmiri	21041	2660	5365	400	29469
Konkani	23144	3000	5744	482	32370
Malayalam	20071	3311	6257	501	30140
Manipuri	10156	2021	3806	332	16351
Marathi	23271	3146	5269	539	32226
Nepali	6748	1477	3227	261	11713
Odiya	27216	2418	5273	377	35284
Punjabi	23255	2836	5830	443	32364
Sanskrit	32385	1246	4006	265	37907
Tamil	16312	2803	5827	477	25419
Telugu	12078	2795	5776	442	21091
Urdu	22990	2801	5786	443	34280

Table 1: Number of synsets in different wordnets

	Nouns		Verbs		Adjectives		Adverbs		Total
	D	H	D	H	D	H	D	H	
Assamese	7019	679	1300	36	2744	0	294	0	12072
Bengali	11049	7680	1824	99	3356	3	312	0	24323
Bodo	6940	603	1594	64	2854	1	293	0	12349
Gujarati	10910	7533	1825	99	3356	3	312	0	24038
Hindi	11584	8221	1988	212	3542	4	344	0	25895
Kannada	7806	1973	1921	154	3453	3	133	0	15443
Kashmiri	9363	6261	1767	100	3240	2	294	0	21027
Konkani	10545	6952	1888	128	3391	2	328	0	23234
Malayalam	9146	4754	1970	206	3525	4	340	0	19945
Manipuri	7192	823	1324	43	2712	0	244	0	12338
Marathi	9874	6556	1839	144	3092	0	333	0	21838
Nepali	5217	496	1114	42	2202	1	200	0	9272
Odiya	11039	7680	1679	66	3187	2	271	0	23924
Punjabi	10215	6382	1822	99	3355	3	312	0	22188
Sanskrit	8396	6470	1048	28	2873	2	241	0	19058
Tamil	8130	3066	1821	98	3353	3	312	0	16783
Telugu	6944	1843	1819	98	3350	0	312	0	14366
Urdu	10424	6816	1822	98	3356	3	313	0	22832

Table 2: Linkage Statistics for English to Indian Language WordNets. D stands for Direct links, and H stands for Hypernymy links

most commonly spoken language in India. The expansion approach adopted for IndoWordNet creation is:

1. Creation of a Hindi synset with synonymous words.
2. Mapping of the synset with relations such as hypernymy and hyponymy *etc.*
3. Tagging of the synset with an ontological category.
4. Allotment of a unique synset ID to the concept de-

scribed in the synset.

5. Creation of the same synset in the other Indian languages leading to an implicit linkage of relations, ontological categories.

We release the latest data in IndoWordNet with statistics described in subsection 3.1.2. below.

3.1.1. Construction Principles

- **Minimality:** We try to capture the minimal set of words in the synset which uniquely define the concept and ensure that it is identifiable via the use of these words.
- **Coverage:** We also try to stress on the completion of the synset and try to capture all the words which represent the concept.
- **Replaceability:** This principle states that all the words in the synset should be able to replace one another in an example sentence quoted along with the synset. These words must be able to replace each other in the same sense.

3.1.2. Current Statistics: IndoWordnet

Table 1 shows the statistics of the released wordnets. These wordnets have, on an average, approximately 28,000 synsets, with Nepali and Hindi having the minimum and the maximum number of synsets respectively. The number of synsets in Hindi is maximum due to the fact that work on IndoWordNet started with the Hindi language. It should also be noted that the ratio of nouns, verbs, adjectives, and adverbs is also on an average 48:6:13:1; the trend being similar to Princeton WordNet.

3.2. Linkage between English and Indian Language WordNets

For linking Indian language wordnets with the Princeton WordNet, we link the Hindi Wordnet data with Princeton WordNet data manually with the help of lexicographers. This has been an ongoing work since many years, and a resource release was long standing. We delve deep into the language related issues in linking both the languages and ensure that only a valid relation is established between both the lexicons. The principles used and the current linkage statistics are described in the subsections below.

3.2.1. Principles

We use the simple principles of concept representation to ensure a valid linkage between the two languages. While linking two concepts, we refer to all words present in both the synsets for creating the linkage. First, we start with linking the known common concepts between both the WordNets of Hindi and English (Direct Linkages). We, then, start to link Hypernymy linkages from Hindi to English. For *e.g., younger paternal uncle* and *elder paternal uncle* are two different specific concepts, and thus have two different synsets in Hindi language. English language, on the other hand, has only the concept of *uncle*, and hence we link both the Hindi language concepts to uncle as Hypernymy linkages.

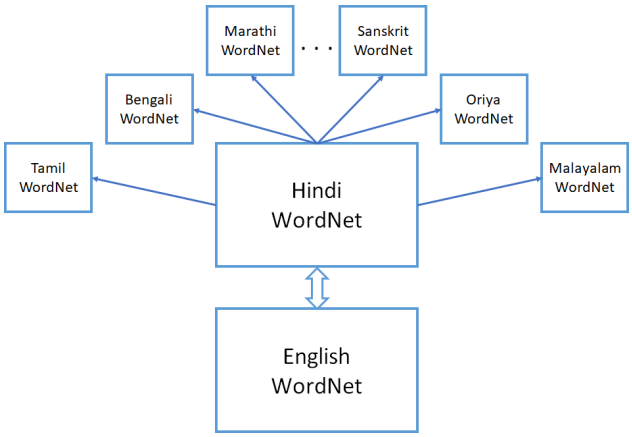

Figure 1: Indian Language WordNet linkages with Princeton WordNet. D stands for links of the type Direct, whereas H stands for the links of the type HYPERNYM.

3.2.2. Princeton Statistics

At present, the Princeton Wordnet has a total of 117659 synsets, with 82115 nouns, 13767 verbs, 18156 adjectives, and 3621 adverbs[3]. They further categorize some of their adjectives into satellite adjectives but the statistics shown include both adjectives and satellite adjectives. We use Princeton WordNet version 3.0 for the purpose of linkage. We began linking Hindi WordNet with version 2.1 and shifted to WordNet version 3.0 using the mappings provided[4] by Princeton WordNet.

3.2.3. Current Statistics: Linkages for Language pairs

Table 2 shows the statistics of the released linkages. There are approximately 20,000 links for an English-Indian language pair on average, with Nepali and Hindi having the minimum and the maximum number of links. Again, the number of links in Hindi is maximum due to the fact that work on IndoWordnet started with the Hindi language, and we link Hindi directly with English. At times, the concept present in Hindi is not present in the other Indian languages thus leading to the less number of linkages for the other languages, in some cases. Table 2 show part-of-speech category-wise distribution of the linked synsets, and also indicated the number of directly linked synsets (D) along with the synset linkages which have been marked as hypernymy linkages (H).

The statistics show our progress in updating IndoWordnet as a resource. The relatively large number of linkages also show that the Indian wordnets have matured considerably.

4. Discussion

Many concepts in the Indian languages are specific to the Indian culture. Thus, their corresponding variant is not available in the Princeton WordNet (and is not likely to be included anytime). Thus, one needs to maintain the translation/transliteration of such notions from Indian languages

[3] https://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

[4] https://wordnet.princeton.edu/man/sensemap.5WN.html

to the English language as a separate bilingual mapping [5]. A similar issue arises in case of proper nouns, which should be present in an Indian lexicon but they are not present in Princeton WordNet. They are also handled using bilingual mappings (Singh et al., 2016). Some of the synsets in Indian languages are too fine-grained and have a common representation in the English language. This is why we use the principle of Hypernymy linkages for linking such concepts. We reserve a set of synset id numbers later for language specific concepts and create them to include in these wordnets, individually. These are not linked to the Princeton WordNet and hence are not included in our resource.

5. Conclusion and Future Work

In this paper, we describe two resources released along with this paper. We discussed the Indian language wordnets that are part of the IndoWordNet project. We enlisted the statistics of the latest version, which we provide as a single bundle along with this paper. Next, we described the linkage process for creating English-Indian language links using English-Hindi language links. We then enlisted the statistics of the latest version of this linked data, which is also provided along with this paper.

In future, we plan to continue building the wordnets and increase linkage. We will also investigate semi-automatic linkage tools such as the ones created by Joshi et al. (2012b), *etc.* so that the workload on our lexicographers and researchers can be reduced to a certain extent

6. Acknowledgement

We would like to acknowledge the work done by lexicographers at Center For Indian Language Technology (CFILT), IIT Bombay without which we would not have been able to link Indian Language Wordnets.

7. Bibliographical References

Balamurali, A. (2012). Cross-lingual sentiment analysis for indian languages using linked wordnets.

Bhattacharyya, P. (2010). Indowordnet. In *In Proc. of LREC-10*. Citeseer.

Bojar, O., Diatka, V., Rychlý, P., Straňák, P., Suchomel, V., Tamchyna, A., and Zeman, D. (2014). HindMonoCorp 0.5.

Bouma, G. (2009). Cross-lingual dutch to english alignment using eurowordnet and dutch wikipedia. In *Proceedings of the 4th International Workshop on Ontology Matching, CEUR-WS*, volume 551, pages 224–229. Citeseer.

Clough, P. and Stevenson, M. (2004). Cross-language information retrieval using eurowordnet and word sense disambiguation. In Sharon McDonald et al., editors, *Advances in Information Retrieval*, volume 2997 of *Lecture Notes in Computer Science*, pages 327–337. Springer Berlin Heidelberg.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J. (1998). Indexing with wordnet synsets can improve text retrieval. *arXiv preprint cmp-lg/9808002*.

Hovy, E. (1998). Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC)*, pages 535–542.

Joshi, S., Chatterjee, A., Karra, A. K., and Bhattacharyya, P. U. (2012a). Eating your own cooking: automatically linking wordnet synsets of two languages.

Joshi, S., Chatterjee, A., Karra, K. A., and Bhattacharyya, P. (2012b). Eating your own cooking: Automatically linking wordnet synsets of two languages. In *Proceedings of COLING 2012: Demonstration Papers*, pages 239–246. The COLING 2012 Organizing Committee.

Knight, K. and Luk, S. K. (1994). Building a large-scale knowledge base for machine translation. In *AAAI*, volume 94, pages 773–778.

Lev Finkelstein, Evgeniy Gabrilovich, Y. M. E. R. Z. S. G. W. and Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Trans. Inf. Syst.*, 20(1):116–131, January.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer.

Singh, M., Shukla, R., Jha, J., Kashyap, L., Kanojia, D., and Bhattacharyya, P. (2016). Mapping it differently: A solution to the linking challenges. In *Eighth Global Wordnet Conference*. GWC 2016.

Sinha, M., Reddy, M., and Bhattacharyya, P. (2006). An approach towards construction and application of multilingual indo-wordnet. In *3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea*.

TufiŞ, D., Ion, R., and Ide, N. (2004). Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1312. Association for Computational Linguistics.

Vossen, P. et al. (1997). Eurowordnet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, pages 5–7.

Yang, D. and Powers, D. M. W. (2006). Verb similarity on the taxonomy of wordnet. In *In the 3rd International WordNet Conference (GWC-06), Jeju Island, Korea*.

[5] Since bilingual mappings are not standardized, we do not release them along with our resources