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

Introduction

- Wordnets (Fellbaum, 1998) have been useful in different NLP 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.
- They have both language-specific information and an interlingual index
- Linked Wordnets have found their application in machine translation (Hovy, 1998), cross-lingual information retrieval (Gonzalo et al., 1998), *etc.*
- Creation and maintenance of Wordnets needs expert involvement - time, resources, and knowledge of multiple languages in case of multiple languages

Contributions

- We release the latest version of 18 wordnets under the IndoWordNet project as a single bundle[1].
- Using mappings between Princeton WordNet and Hindi wordnet, we create and release mappings between Princeton WordNet and these 18 languages wordnet.

Background and Related Work

- Princeton WordNet or the English WordNet was the first Wordnet.
- EuroWordNet (Vossen et al., 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 of these wordnets separately capture a language-specific information.
  - These wordnets are also 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 listed in Table 1.
- These wordnets have been created using the expansion approach with Hindi WordNet as a pivot, which is partially linked to English WordNet.

Resources Released

Indian Language WordNets

- The creation of IndoWordNet began in 2000 with Hindi WordNet.
- Hindi was chosen as a pivot as it shares many common features and borrowed concepts from ancient Indian languages like Sanskrit and is the 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 described in the synset.
  5. Creation of the same synset in the other Indian languages leading to an implicit linkage of relations, ontological categories.

Current Statistics: IndoWordnet

- 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.

	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*

Linkages between English and Indian Language WordNets

- Hindi is ***the pivot*** for IndoWordNet.
- If we link Hindi Wordnet with Princeton WordNet, we have linkages between all languages of IndoWordNet and Priceton WordNet.
- This linking is done with the help of lexicographers using the following principles.
  - 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.
    - Example: *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.

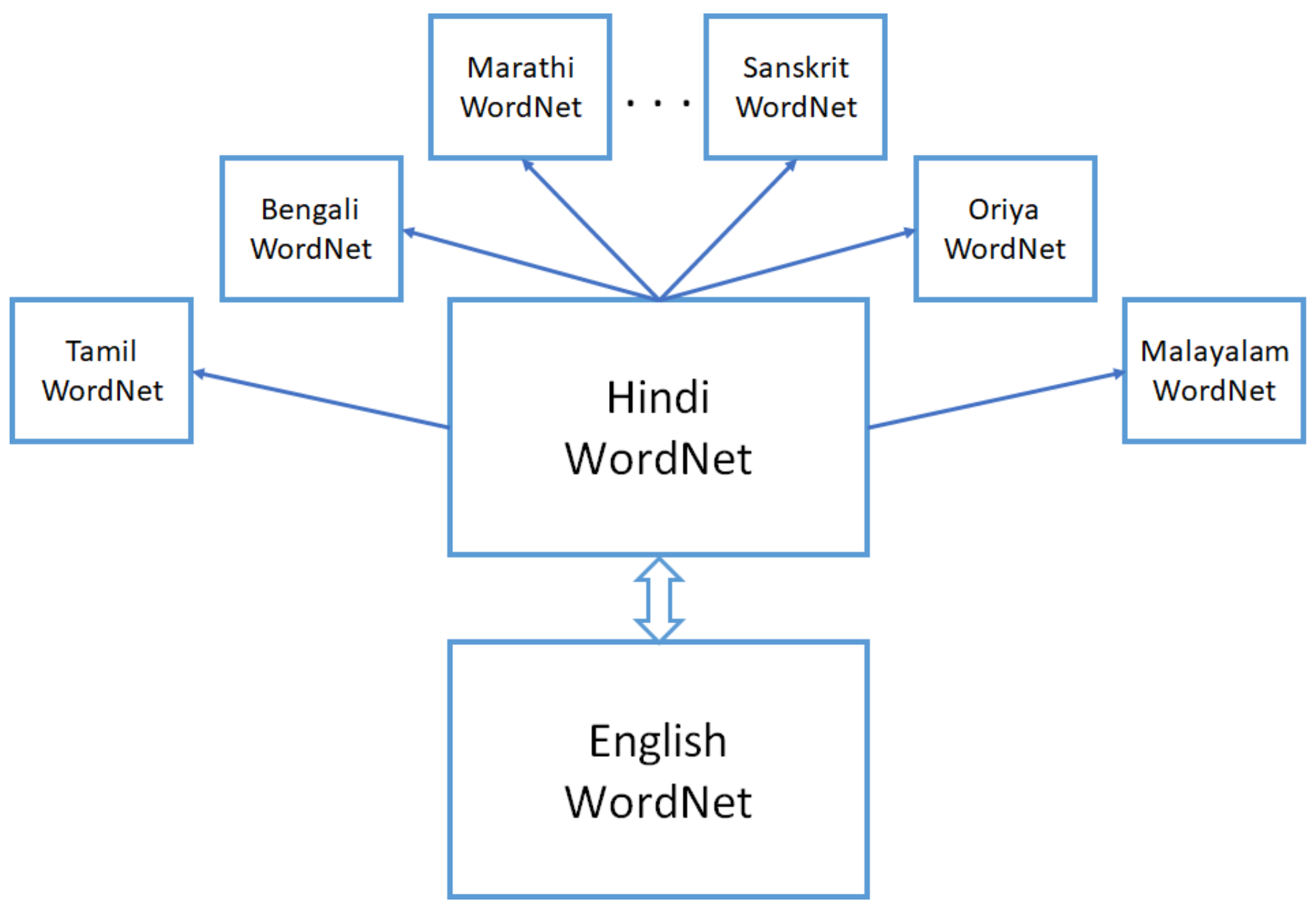

Figure 1: *Indian Language WordNet linkages with Princeton WordNet. D stands for links of the type Direct, whereas H stands for the links of the type HYPERNYM.*

Princeton Statistics

- Princeton Wordnet has a total of 117659 synsets, with 82115 nouns, 13767 verbs, 18156 adjectives (including satellites), and 3621 adverbs[2].
- We use Princeton WordNet version 3.0 for the purpose of linkage.

Current Statistics: Linkages for Language pairs

- 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.
- 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.
- The relatively large number of linkages in the statistics show that Indian Wordnets have matured considerably.
- Translation/Transliteration of those Indian culture-specific concepts whose corresponding concepts are missing in Princeton WordNet, are maintained separately as a separate bilingual mapping.

	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*

Conclusion and Future Work

- We described two resources released with this paper.
- We discussed the Indian language wordnets that are part of the IndoWordNet project and enlisted the statistics of the latest version.
- We described the linkage process for creating English-Indian language links using English-Hindi language links and enlisted the statistics of the latest version of this linked data.
- In future, we plan to continue building the wordnets and increase linkage.

References

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

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.

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

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.

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.

[1] `http://www.cfilt.iitb.ac.in/ilw`
[2] `https://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html`

Presented at the 11th edition of the Language Resources and Evaluation Conference, 7-12 May 2018, Miyazaki (Japan)