The IndoWordnet Parallel Corpus

Anoop Kunchukuttan
anoop.kunchukuttan@gmail.com

1 Summary

IndoWordnet (Bhattacharyya, 2010) is a linked structure of wordnets of major Indian languages from Indo-Aryan, Dravidian and Sino-Tibetan families. Synsets are linked across many languages. Every synset in every language contains a gloss and example usage sentence/phrase. In a large number of cases, the example and gloss sentences across languages are translations. Hence, IndoWordNet is a source of parallel corpora across multiple Indian languages.

We mine two parallel corpora from IndoWordNet using the synset databases provided by the *pyiwn* project (`https://github.com/riteshpanjwani/pyiwn/`) (Panjwani et al., 2018):

- **Gloss:** These are synset definitions. They could be phrase or entire sentences.
- **Examples:** These are sentences depicting usage of one of the words in the synset. Typically, these are complete sentences.

A cursory observation shows that most sentences are translations, though there are some cases where the the gloss/example may be different. Such divergence seems to be more in the case of examples. In the next revision of the corpus, we could explore refinement of the parallel corpora.

2 Corpus Statistics

The corpus contains parallel corpora from the following 18 langauges:

- Indo-Aryan: Assamese, Bengali, Gujarati, Hindi, Kashmiri, Konkani, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Urdu.
- Dravidian: Kannada, Malayalam, Tamil, Telugu.
- Tibeto-Burman: Bodo, Meitei (Manipuri).

The corpus contains around 12.5 million segments across all languages and the two sources (gloss and examples). The following are the overall statistics of the corpus:

Corpus	Sentences/Segments
Gloss	6,297,688
Examples	6,296,223
Total	12,593,911

Table 1: Summary Statistics

Pairwise statistics for each language pair of the Gloss and Example corpora are shown in Table 2 and 3 respectively.

3 Data Format

We distribute two files:

- *gloss.csv*: contains parallel corpora extracted from synset glosses.
- *example.csv*: contains parallel corpora extracted from synset examples.

Each line in both files contains translations of one sentence. The translation in different languages are delimited by three pipe characters (|||). The first row indicates the language of each column.

4 License

This dataset is released under the *Creative Commons Attribution Share Alike 4.0 International* license, the same license as *pyiwn*.

References

Bhattacharyya, P. (2010). IndoWordNet. In *In Proceedings of LREC*.

Panjwani, R., Kanojia, D., and Bhattacharyya, P. (2018). pyiwn: A Python-based API to access Indian Language WordNets. In *Proceedings of the Global WordNet Conference*, volume 2018.

Table 2: Number of sentence pairs in the IndoWordnet Gloss parallel corpus

	asm	ben	brx	guj	hin	kan	kas	kok	mal	mar	mni	nep	ori	pan	san	tam	tel	urd
asm		14948	14106	1492	14618	13166	1411	14938	14854	14156	14717	10931	14763	14584	11986	14187	13137	14849
ben	14948		15773	35481	35666	2049	28742	31253	28745	27177	1554	1168	35179	32357	23722	25394	21085	34232
brx	14106	15773		15742	15423	13914	15185	15752	15678	14595	14301	11515	15455	15404	11671	15055	14024	15681
guj	1492	35482	15742		34949	20453	28375	30933	28436	26962	15495	11662	34446	3176	23549	25185	20979	33562
hin	14618	35666	15423	34948		21786	28926	31651	29561	29142	15214	11432	34667	31897	25055	25015	20742	33773
kan	13166	2049	13914	20453	21786		18597	20695	20512	18977	13401	10363	19541	19413	1552	19103	18786	19651
kas	14127	28776	15209	28407	28959	18623		26494	24618	24006	14561	11314	27924	2646	20174	21609	18132	27449
kok	14938	31253	15752	30932	31651	20695	2646		25523	26889	15395	1166	30375	28366	22425	22325	19477	2954
mal	14854	28745	15678	28436	29561	20512	24584	25523		23308	15334	11668	27713	27524	19158	25138	20979	28041
mar	14156	27177	14595	26961	29142	18977	23976	26889	23308		14466	10922	26478	25071	20955	20101	17524	25803
mni	14717	1554	14301	15495	15214	13401	14542	15395	15334	14466		11244	15454	15136	12178	14595	13418	15411
nep	10931	1168	11515	11662	11432	10363	11303	1166	11668	10922	11244		11687	11371	8891	11396	10547	11617
ori	14763	35179	15455	34445	34667	19541	27898	30375	27713	26478	15454	11687		31306	23253	24359	20035	3318
pan	14584	32357	15404	31759	31897	19413	26427	28366	27524	25071	15136	11371	31306		20921	24557	20292	31148
san	11995	23739	11679	23565	25071	15531	20173	22442	19173	20967	12187	8897	23272	20934		16342	14822	22254
tam	14189	25397	15057	25188	25018	19106	21585	22328	25141	20104	14597	11398	24362	2456	16331		21071	25341
tel	13137	21085	14024	20979	20742	18786	18108	19477	20979	17524	13418	10547	20035	20292	1481	21068		21046
urd	14849	34232	15681	33561	33773	19651	27417	2954	28041	25803	15411	11617	3318	31148	22239	25338	21046	

Table 3: Number of sentence pairs in the IndoWordnet Example parallel corpus

	asm	ben	brx	guj	hin	kan	kas	kok	mal	mar	mni	nep	ori	pan	san	tam	tel	urd
asm		14948	14106	14917	14618	13165	14108	14936	14854	14153	14713	10931	14762	14584	1198	14163	13137	14843
ben	14944		15773	35471	35666	20488	28739	31246	28745	27174	15536	1168	35178	32357	23713	25339	21085	34224
brx	14103	15773		15738	15423	13913	15184	1575	15678	14593	14297	11515	15454	15404	11664	15032	14024	15675
guj	14916	35482	15742		34949	20451	28371	30926	28436	26959	15491	11662	34445	3176	2354	25131	20979	33554
hin	14614	35666	15423	34938		21784	28922	31632	29561	29139	1521	11432	34666	31897	25046	2496	20742	33766
kan	13164	2049	13914	20442	21786		18593	20678	20512	18974	13398	10363	19541	19413	15513	19062	18786	19645
kas	14124	28776	15209	2840	28959	18621		26489	24618	24003	14557	11314	27924	2646	20165	21578	18132	27443
kok	14934	31253	15752	30923	31651	20694	26459		25523	26886	15391	1166	30374	28366	22417	22279	19477	29532
mal	1485	28745	15678	28425	29561	2051	24582	25511		23305	1533	11668	27712	27524	19149	25085	20979	28035
mar	14152	27177	14595	26959	29142	18976	23979	26873	23308		14462	10922	26477	25071	20948	20055	17524	25797
mni	14713	1554	14301	1549	15214	1340	14541	15393	15334	14463		11244	15453	15136	12172	14572	13418	15405
nep	10929	1168	11515	11658	11432	10362	11299	11659	11668	1092	1124		11687	11371	8885	11386	10547	11611
ori	14759	35179	15455	34435	34667	1954	27894	30369	27713	26476	1545	11687		31306	23244	24304	20035	33172
pan	1458	32357	15404	3175	31897	19411	26425	28359	27524	25068	15132	11371	31305		20913	24505	20292	3114
san	11993	23739	11679	23561	25069	15529	20165	22437	19173	20965	12184	8897	23272	20934		16311	14822	22248
tam	14186	25397	15057	25177	25018	19104	2158	22323	25141	20101	14593	11398	24361	2456	16323		21071	25335
tel	13135	21085	14024	20968	20742	18784	18103	19473	20979	17521	13415	10547	20035	20292	14802	21019		2104
urd	14845	34232	15681	33551	33773	19649	27413	29534	28041	2580	15407	11617	33179	31148	2223	25283	21046	