The IndoWordnet Parallel Corpus

Anoop Kunchukuttan
anoop.kunchukuttan@gmail.com

14 May 2020

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.

Corpus Website: `https://github.com/anoopkunchukuttan/indowordnet_parallel`
Version: v0.2

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

Corpus	Parallel Sentences/Segments
Gloss	3,148,483
Examples	3,147,020
Total	6,295,503

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	14920	14618	13166	14110	14938	14854	14156	14717	10931	14763	14584	11986	14187	13137	14849
ben	14948		15773	35481	35666	20490	28742	31253	28745	27177	15540	11680	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	14920	35481	15742		34948	20453	28374	30932	28436	26961	15495	11662	34445	31759	23548	25185	20979	33561
hin	14618	35666	15423	34948		21786	28926	31651	29561	29142	15214	11432	34667	31897	25055	25015	20742	33773
kan	13166	20490	13914	20453	21786		18597	20695	20512	18977	13401	10363	19541	19413	15520	19103	18786	19651
kas	14110	28742	15185	28374	28926	18597		26460	24584	23976	14542	11303	27898	26427	20163	21582	18108	27417
kok	14938	31253	15752	30932	31651	20695	26460		25523	26889	15395	11660	30375	28366	22425	22325	19477	29540
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	15540	14301	15495	15214	13401	14542	15395	15334	14466		11244	15454	15136	12178	14595	13418	15411
nep	10931	11680	11515	11662	11432	10363	11303	11660	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	33180
pan	14584	32357	15404	31759	31897	19413	26427	28366	27524	25071	15136	11371	31306		20921	24557	20292	31148
san	11986	23722	11671	23548	25055	15520	20163	22425	19158	20955	12178	8891	23253	20921		16329	14810	22239
tam	14187	25394	15055	25185	25015	19103	21582	22325	25138	20101	14595	11396	24359	24557	16329		21068	25338
tel	13137	21085	14024	20979	20742	18786	18108	19477	20979	17524	13418	10547	20035	20292	14810	21068		21046
urd	14849	34232	15681	33561	33773	19651	27417	29540	28041	25803	15411	11617	33180	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		14944	14103	14913	14614	13163	14105	14932	14850	14149	14709	10929	14759	14580	11978	14160	13135	14839
ben	14944		15773	35471	35666	20488	28739	31246	28745	27174	15536	11680	35178	32357	23713	25339	21085	34224
brx	14103	15773		15738	15423	13913	15184	15750	15678	14593	14297	11515	15454	15404	11664	15032	14024	15675
guj	14913	35471	15738		34938	20440	28363	30916	28425	26956	15486	11658	34434	31750	23535	25120	20968	33543
hin	14614	35666	15423	34938		21784	28922	31632	29561	29139	15210	11432	34666	31897	25044	24960	20742	33766
kan	13163	20488	13913	20440	21784		18591	20677	20510	18973	13397	10362	19540	19411	15511	19060	18784	19643
kas	14105	28739	15184	28363	28922	18591		26454	24582	23976	14537	11299	27894	26425	20146	21546	18103	27407
kok	14932	31246	15750	30916	31632	20677	26454		25511	26870	15389	11659	30368	28359	22412	22274	19473	29526
mal	14850	28745	15678	28425	29561	20510	24582	25511		23305	15330	11668	27712	27524	19149	25085	20979	28035
mar	14149	27174	14593	26956	29139	18973	23976	26870	23305		14459	10920	26475	25068	20946	20052	17521	25794
mni	14709	15536	14297	15486	15210	13397	14537	15389	15330	14459		11240	15449	15132	12169	14568	13415	15401
nep	10929	11680	11515	11658	11432	10362	11299	11659	11668	10920	11240		11687	11371	8885	11386	10547	11611
ori	14759	35178	15454	34434	34666	19540	27894	30368	27712	26475	15449	11687		31305	23244	24303	20035	33171
pan	14580	32357	15404	31750	31897	19411	26425	28359	27524	25068	15132	11371	31305		20913	24505	20292	31140
san	11978	23713	11664	23535	25044	15511	20146	22412	19149	20946	12169	8885	23244	20913		16291	14802	22224
tam	14160	25339	15032	25120	24960	19060	21546	22274	25085	20052	14568	11386	24303	24505	16291		21019	25277
tel	13135	21085	14024	20968	20742	18784	18103	19473	20979	17521	13415	10547	20035	20292	14802	21019		21040
urd	14839	34224	15675	33543	33766	19643	27407	29526	28035	25794	15401	11611	33171	31140	22224	25277	21040	