Samayik: A Benchmark and Dataset for English-Sanskrit translation

Ayush Maheshwari, Ashim Gupta, Amrith Krishna, Atul Kumar Singh
Ganesh Ramakrishnan, G. Anil Kumar, Jitin Singla

Introduction

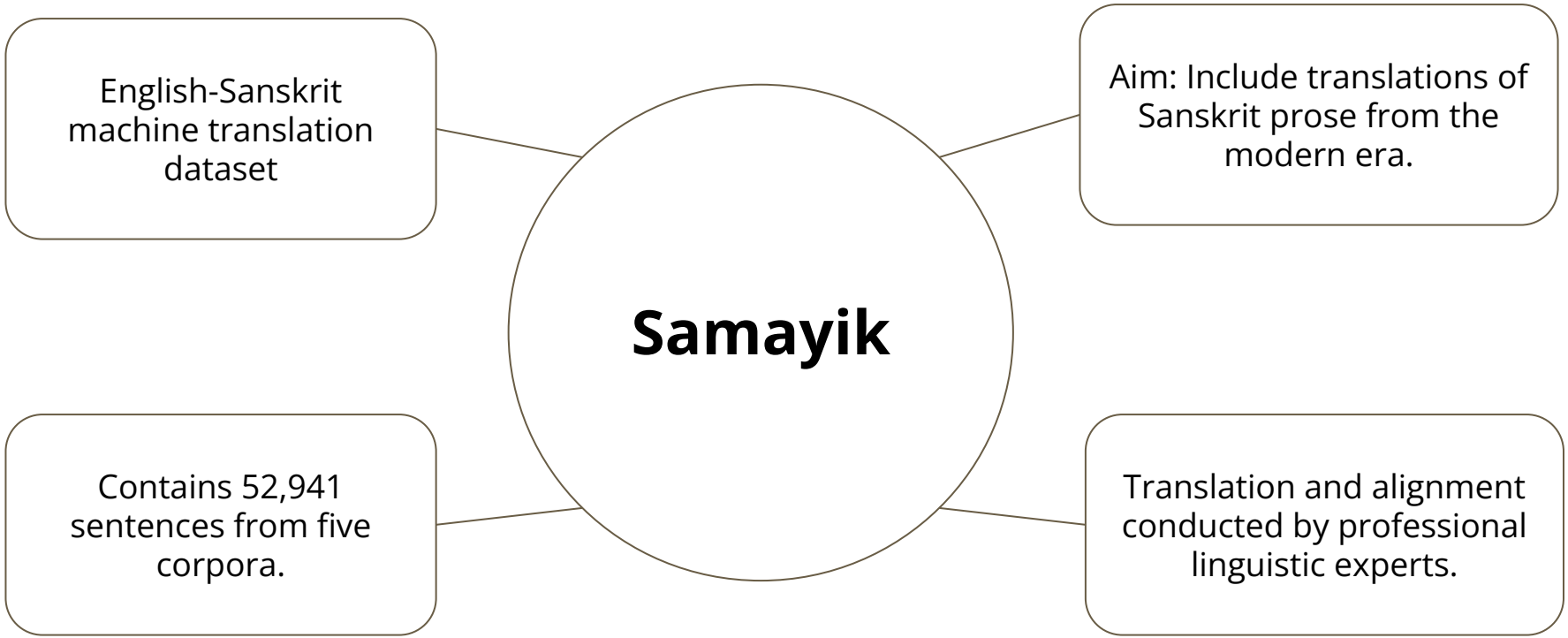
Sanskrit, renowned for its classical heritage, encounters digitization challenges, being labeled a low-resource language with just one million monolingual sentences available digitally.

Sanskrit possesses approximately 30 million extant manuscripts suitable for digitization. It boasts over two million active speakers.

Sentence constructions in Sanskrit exhibit relatively free word order, with verse form adhering to prescribed meter patterns and prose generally following Subject-Object-Verb (SOV) ordering.

To address this gap, we introduce Sāmayik.

Samayik:


English-Sanskrit machine translation dataset
Aim: Include translations of Sanskrit prose from the modern era.
Samayik
Contains 52,941 sentences from five corpora.
Translation and alignment conducted by professional linguistic experts.

Samayik dataset resources

Samayik

Bible - The New Testament:

- Utilized Sanskrit version by Calcutta Baptist Missionaries (1851).
- New Testament comprises 7,838 sentences from 260 chapters.
- One-to-one correspondence at the sentence level for both English and Sanskrit sentences.
- Alignment verification by three fluent speakers for 100 sentences.

Samayik dataset resources

Samayik

Mann ki Baat (MKB):

- Ongoing monthly radio podcast hosted by the Prime Minister of India.
- Addresses social, cultural, and contemporary topics.
- Sanskrit translations available in public domain.
- Manually aligned Sanskrit sentences with official English transcripts from 25 episodes.
- Further verification by in-house language experts.
- MKB English-Sanskrit corpus "Sāmayik" release: 4,047 sentences; 47,843 words.

Samayik dataset resources

Samayik

Gītā Sopānaṁ:

- Book by ‘Samskrita Bharati’ (2009) for teaching Sanskrit to beginners.
- Contains 6,130 sentences with 6,465 unique words.
- Focuses on learning grammar through stories.
- In-house translation to English sentences by 4 language experts.
- Expert-level annotations with one translation per Sanskrit sentence.

Samayik dataset resources

Samayik

Spoken Tutorials:

- Large corpus of video tutorials for open-source software training.
- Translated into several languages, including Sanskrit.
- Transcripts of 254 videos extracted, each approximately 10 minutes long.
- Alignment between English and corresponding Sanskrit sentences performed manually by 5 linguistic experts.
- Final corpus comprises 23,835 sentences; 237,449 words.

Samayik dataset resources

Samayik

NIOS (National Institute of Open Schooling):

- National-level board of education established in 1989.
- Provides self-instructional study materials for various subjects.
- Obtained study materials from Indian knowledge tradition courses offered by NIOS.
- Conversion of PDF files to text format using PDF parsers.
- Alignment of sentences by a team of five English and Sanskrit linguistic experts.
- NIOS corpus contains 11,356 parallel sentences; 105,178 total words; 30,966 unique words.

Dataset stats

Dataset	NIOS	Spoken Tutorials	GitaSopanam	Bible	Mann Ki Baat	Total
#sentences	11356	23835	5885	7838	4047	52941
#words	105178	237449	26135	102508	47843	518842
#unique words	30966	38373	6513	38359	20484	122349
% of unique words	29.4	16.2	24.9	37.4	42.8	23.6
Mean word length	9.3	10	4.5	13.1	11.8	9.8

Experimental setup

We fine-tuned 4 different models that Sanskrit words in their vocabulary to check the information gained by the model using our data. These are:

mBart: A multilingual pretrained seq2seq model, mbart-large-50-many-to-many-mmt, trained on a large corpus of 50 languages

IndicBart: A multilingual pretrained seq2seq model with 244M parameters, trained on corpora from Indic languages and English.

ByT5: A multilingual pretrained seq2seq model with 244M parameters, trained on corpora from Indic languages and English

Indictrans2: A multi-lingual translation model trained on 22 Indic languages, including Sanskrit.

Test result in In-domain data

	En-Sa		Sa-En	
Model	**BLEU**	**ChrF**	**BLEU**	**ChrF**
ByT5	**28.7**	44.4	31.1	55.7
mBART	27.20	**46.61**	11.6	27.0
IndicBART	25.45	43.47	29.79	50.14
Indictrans	11.3	46.6	**37**	**58.2**

Results for different models on the indomain (corpora NIOS, Spoken Tutorial , Gita Sopanam , and Bible test set for En-Sa and Sa-En direction.

Test result in out of domain data

	En-Sa		Sa-En	
Model	**BLEU**	**ChrF**	**BLEU**	**ChrF**
ByT5	7	21.4	5.4	29
mBART	**7.11**	22.6	-	-
IndicBART	6.9	22.4	5.3	27.7
Indictrans	0.6	**26.7**	13.1	37.5
Google Trans	1.9	35	**13.9**	**44.7**
NLLB	1.2	27.6	11.5	36.1
Indictrans(Vanilla)	1.2	34	14.5	42.7

Results for out-of-domain test set, namely,
Mann Ki Baat (MKB) for En-Sa and Sa-En directions.

Conclusion

1. **Dataset Introduction:**

   - We unveil Samayik, a novel dataset featuring 52,000+ sentences designed for English-Sanskrit translation.

   - Samayik diverges from prior datasets by focusing on contemporary prose across various domains like instructional material and radio podcasts.

2. **Baseline Models:**

   - We provide a suite of robust baselines constructed on four multilingual pretrained models.

3. **Performance Evaluation:**

   - Empirical assessments reveal that models trained on our dataset outperform those trained on existing datasets.

   - Furthermore, our models exhibit superior performance compared to pretrained models enriched with a Sanskrit corpus.