Sāmayik: A Benchmark and Dataset for English-Sanskrit Translation

Ayush Maheshwari[1], **Ashim Gupta**[2], **Amrith Krishna**[3], **Ganesh Ramakrishnan**[1], **G. Anil Kumar**[4], **Jitin Singla**[5]
[1]Indian Institute of Technology Bombay
[2]University of Utah
[3]Uniphore Inc.
[4]Dept. of Physics, Indian Institute of Technology Roorkee
[4]Dept. of Biosciences and Bioengineering, Indian Institute of Technology Roorkee

arXiv:2305.14004v1 [cs.CL] 23 May 2023

Abstract

Sanskrit is a low-resource language with a rich heritage. Digitized Sanskrit corpora reflective of the contemporary usage of Sanskrit, specifically that too in prose, is heavily under-represented at present. Presently, no such English-Sanskrit parallel dataset is publicly available. We release a dataset, Sāmayik of more than 42,000 parallel English-Sanskrit sentences, from four different corpora that aim to bridge this gap. Moreover, we also release benchmarks adapted from existing multilingual pretrained models for Sanskrit-English translation. We include training splits from our contemporary dataset and the Sanskrit-English parallel sentences from the training split of Itihāsa, a previously released classical era machine translation dataset containing Sanskrit.

1 Introduction

Sanskrit is a classical language, with a rich heritage spanning more than three millennia. Moreover, it is a language in sustenance with more than two million active speakers (McCartney, 2019; Chandramouli, 2011). While Sanskrit is a heritage-rich language, it still is considered a low-resource language (Hellwig, 2010–2021; Maheshwari et al., 2022). Moreover, the available corpora often cover content that is vastly divergent in terms of their domains, chronology, stylistic features, usage, syntactic features (Hellwig, 2009), and even in their typological features such as word-order (Krishna et al., 2021; Tubb and Boose, 2007). In this work, we release a parallel Sanskrit-English dataset that covers multiple corpora representing contemporary Sanskrit.

'Itihāsa' currently forms the largest parallel machine translation corpus containing Sanskrit as one of the languages (Aralikatte et al., 2021). It is a Sanskrit-English dataset, containing 93,000 pairs of verses in Sanskrit and their corresponding translation in English. These sentences were collected from two epics written in the classical era with their translations sometime in the early half of the twentieth century. Similarly, the Digital Corpus of Sanskrit (DCS) currently forms the largest monolingual dataset in Sanskrit (Hellwig, 2010–2021). DCS contains more than 600,000 monolingual sentences in Sanskrit, spanning a chronology of around 2000 years categorized into pre-classical literature (1500 BCE - 100 BCE), classical literature (300 CE - 800 CE), and modern literature (900 CE to now; Krishna et al., 2018). However, currently, available digitized content in modern literature also is mostly confined to that written until the first half of the twentieth century.

Identifying a gap in Sanskrit sentences representing contemporary Sanskrit, mostly focusing on content written in the second half of the twentieth century to now, we release our dataset *Sāmayik*[1]. The dataset is a Sanskrit term that translates to the "sayings of the contemporary world". Sāmayik consists of 43,000 parallel sentence pairs, collected from four different sources. These are spoken content that covers contemporary world affairs, interpretation of literary works, pedagogical content, *etc*. In Table 1, we provide statistics about each of the individual corpus and that of our complete dataset. We describe each corpus in Section 2. The oldest corpus in our collection is the English-Sanskrit Bible, where the Sanskrit translation was performed in 1851 and it forms less than 20 % of the overall dataset. The Sanskrit component in the rest of the corpora is either created in the latter half of the twentieth century or in the current century. The latest corpus in our collection contain content as latest as 2022, from Sanskrit and English transcriptions of 'Mann Ki Baat' a podcast currently in production.

Sanskrit is a morphologically rich language and is lexically productive. Moreover, sentence constructions in Sanskrit follow relatively free word

[1]The data will be released after publication.

order. Here, sentences written in verse form have to adhere to prescribed meter patterns as per prosody. Hence, word order need not adhere to a fixed word-order pattern. However, sentences written in prose tend to form Subject-Object-Verb (SOV) ordering. While Ithihāsa consists of two epics written in poetry form, DCS consists of most of its content in poetry. On the contrary, our corpus focuses on sentences written in prose form.

In addition to our dataset, we release benchmarks by adapting pre-trained models for neural machine translation in Sanskrit-English. Currently, there exist no pre-trained models that include Sanskrit in their benchmarks. Hence, we adapt three pre-trained multilingual seq2seq models for the task, namely ByT5 (Xue et al., 2022), mBART (Liu et al., 2020), and IndicBART (Dabre et al., 2022). IndicBART is a pre-trained model fine-tuned specifically for several Indic languages and English. Further, all the Indic languages are transliterated into Devanagari script, widely used for Sanskrit as well. Similarly, ByT5 is a token-free model which tokenizes inputs at the Unicode byte level and the Devanagari script for Sanskrit is part of the Unicode specifications.

With both MBART and IndicBART, we observe negligible OOV vocabulary subword tokens, and we observe that IndicBART currently reports the best BLEU score on our dataset with a BLEU score of 27.25. Further, we include the Ithihāsa's training split in our training data for comprehensiveness. Additionally, we explore the utility of Hindi as a bridge language in training NMT models for English-Sanskrit. Here, we utilize a subset of the parallel sentences between all three languages, an auxiliary set of Sanskrit-Hindi pairs, and additional publicly available Hindi-Sanskrit pairs.

2 Sāmayik

Sāmayik is an English-Sanskrit machine translation dataset, consisting of 42,961 sentences from four different corpora. The primary aim of the dataset is to include translation pairs containing Sanskrit prose written in the modern era. Here, we give a brief description of each of the datasets involved and the steps involved in processing these sentences. All the sentences in Sanskrit are aligned with their corresponding parallel sentence(s) in English.

Bible - The New Testament: We release the new testament of the Bible aligned with its corresponding English version. We use the Sanskrit version released by Calcutta Baptist Missionaries, originally published in 1851[2]. The new testament essentially contains 7,840 sentences from 260 chapters. Each verse is generally indexed by the Book-name, chapter name followed by the verse number. For the English version of the Bible, we rely on Christodouloupoulos and Steedman (2015) where the English sentences also follow the same indexing form. Given the one-to-one correspondences at the sentence level for both English and Sanskrit sentences, the mapping was straightforward. We finally obtain a total of 7840 parallel sentences. Further, three fluent speakers of both English and Sanskrit have verified the alignments for a sample of 100 sentences.

Mann ki Baat (MKB)[3] - MKB is a monthly Indian radio program hosted by the Prime Minister of India originally in the Hindi language from 2014-2022. Each episode is an address to the nation discussing social, cultural, and contemporary topics including conversation with individuals. The official translations of the transcripts are present in several Indian languages except Sanskrit. However, unofficial Sanskrit translation by experts are available in public domain[4]. We use these expert translations and manually align Sanskrit sentences with official English transcripts from the 25 episodes. Additionally, these Sanskrit translations are further verified by in-house language experts. The MKB corpus consists of 4061 sentences with a total of 47,838 words.

Gītā Sopānaṁ - We extract sentences from the Sanskrit learning book, Gītā Sopānaṁ published in 2009. We ask language experts, well versed in both English and Sanskrit, to translate these sentences into English. Gītā Sopānaṁ is a self-learning book to learn Sanskrit through stories. It often contains simple and small sentences with a focus on learning grammar instead of expanding vocabulary. Therefore, the dataset contains 6465 unique words for a total of 6130 sentences.

[2] https://www.bible.com/bible/2104/MAT.1.SAN-DN

[3] https://pmonradio.nic.in/

[4] https://sanskritdocuments.org/sites/manogatam/

Dataset	Mann Ki Baat	Spoken Tutorials	GitaSopanam	Bible	Total
#sentences	4061	24930	6130	7840	42961
#words	47838	245666	26581	102526	422770
#unique words	19761	38349	6465	37193	95838

Table 1: Statistics for different corpus in the Sāmayik.

Spoken Tutorials[5] - Spoken Tutorial project is a large corpus of video tutorials for training students to use open source software. These tutorials are created by domain experts and translated into several languages by language experts. We scraped[6] videos and transcripts from their website for which both English and the corresponding Sanskrit translations are available. We extracted transcripts of 254 videos where each video is of average 10 minutes in duration. We deploy experts having knowledge of both English and Sanskrit to manually align transcripts from each video. The final corpus contains 24,930 sentences comprising 245,666 words.

3 Preliminary Experiments

3.1 Systems

mBART (Liu et al., 2020): is a multilingual pretrained seq2seq model trained using similar objective as employed in BART (Lewis et al., 2020). We employ mBART-50, trained on a large multilingual corpora of 50 languages. In our experiments, we observe use of SLP1 encoding for Sanskrit leads to the best results.

IndicBART (Dabre et al., 2022) is also a multilingual pretrained seq2seq model following the pretraining objective of BART. However, here the corpora used are specifically from Indic languages and ENglish. While different Indic languages use different scripts, these are losslessly converted to Devanagari before tokenisation during its pretraining. Hence, we use the Devanagari script for encoding Sanskrit, and English uses its roman script.

ByT5 (Xue et al., 2022) is a token free pretrained seq2seq model following pretraining objective as that of T5, or more specifically mT5. However, here it is a token free model that uses a fixed 256 byte values in Unicode as its vocabulary.

[5] https://spoken-tutorial.org/

[6] The website content is licensed under CC4.0 license.

Model	BLEU	ChrF
mBART	19.4	33.2
ByT5	18.8	29.5
IndicBART	**27.3**	**45.7**

Table 2: BLEU and ChrF scores for the test set on English-Sanskrit translation for different pre-trained models.

3.2 Metrics

We evaluate these models on both BLEU and ChrF. BLEU is a word-level n-gram precision-based metric whereas ChrF is a character-level ngram F-score. Here, given that Sanskrit is a morphologically rich language with more than 1,400 possible inflected forms (Krishna et al., 2021), we believe ChrF can be indicative of capturing morpho-syntactic aspects.

3.3 Results

We split our dataset comprising of 42k sentences into 80% for the train set and rest for the evaluation set. The evaluation set is equally split into development and test set. We performed preliminary experiments using pre-trained BART models mBART, ByT5 and IndicBart. We fine-tune over these pre-trained models for English-Sanskrit translation.

Implementation Details All models are trained using HuggingFace Transformers (Wolf et al., 2020). Both source and target sequences are truncated at 512 token length. We use respective model pre-trained tokenizers on our dataset. We use batch size of 128, and use standard cross entropy loss with label smoothing of 0.1 and AdamW optimizer(Loshchilov and Hutter). The model is trained for a maximum of 30 epochs with a learning rate of 1e-3 and weight decay of 1e-4. In order to accommodate bigger models (byT5 and mBART) into memory, we introduce gradient accumulation and increasing the number of epochs to maintain effective batch size and optimization steps.

In Table 2, we present test scores on Sanskrit-English translation. We observe that IndicBART achieves far better BLEU and ChrF scores than mBART and ByT5. This suggests that models trained on Indic language data demonstrates more closeness to the new Indic language from the same language family than models trained on mixed language family.

References

Rahul Aralikatte, Miryam de Lhoneux, Anoop Kunchukuttan, and Anders Søgaard. 2021. Itihasa: A large-scale corpus for Sanskrit to English translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 191–197, Online. Association for Computational Linguistics.

C Chandramouli. 2011. Census of india 2011. *Provisional Population Totals. New Delhi: Government of India*.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49:375–395.

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, and Pratyush Kumar. 2022. IndicBART: A pre-trained model for indic natural language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Oliver Hellwig. 2009. Extracting dependency trees from sanskrit texts. In *Sanskrit Computational Linguistics: Third International Symposium, Hyderabad, India, January 15-17, 2009. Proceedings*, pages 106–115. Springer.

Oliver Hellwig. 2010–2021. Dcs - the digital corpus of sanskrit.

Amrith Krishna, Bishal Santra, Sasi Prasanth Bandaru, Gaurav Sahu, Vishnu Dutt Sharma, Pavankumar Satuluri, and Pawan Goyal. 2018. Free as in free word order: An energy based model for word segmentation and morphological tagging in Sanskrit. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2550–2561, Brussels, Belgium. Association for Computational Linguistics.

Amrith Krishna, Bishal Santra, Ashim Gupta, Pavankumar Satuluri, and Pawan Goyal. 2021. A Graph-Based Framework for Structured Prediction Tasks in Sanskrit. *Computational Linguistics*, 46(4):785–845.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ayush Maheshwari, Nikhil Singh, Amrith Krishna, and Ganesh Ramakrishnan. 2022. A benchmark and dataset for post-OCR text correction in Sanskrit. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6258–6265, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick McCartney. 2019. Sustainably–speaking yoga: Comparing sanskrit in the 2001 and 2011 indian censuses. In *The GLOCAL in Asia 2019*. The GLOCAL Unit, SOAS University of London.

Gary A Tubb and Emery R Boose. 2007. *Scholastic Sanskrit*. American Institute of Buddhist Studies.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.