Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sense-Augmented Corpus for Sanskrit-English Machine Translation: Corpus Construction, Model Fine-Tuning, and Evaluation

Anagha Pradeep

anagha.pradeep@research.iiit.ac.in

IIIT Hyderabad

Sriram Krishnan

Central Sanskrit University

Ketaki Shetye

IIIT Hyderabad

Bassam Adnan

IIIT Hyderabad

Radhika Mamidi

IIIT Hyderabad

Research Article

Keywords: Sense-Augmentation, Machine Translation, Word Sense Disambiguation (WSD), Corpus Construction

Posted Date: September 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10759503/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Sense-Augmented Corpus for Sanskrit-English Machine Translation: Corpus Construction, Model Fine-Tuning, and Evaluation

Anagha Pradeep[1*], Sriram Krishnan[2], Ketaki Shetye[1], Bassam Adnan[1], Radhika Mamidi[1]

[1*]Language Technology Research Centre, IIIT Hyderabad, Professor CR Rao Road, Hyderabad, 500032, Telangana, India.
[2]Central Sanskrit University, Nashik, 422003, Maharashtra, India.

*Corresponding author(s). E-mail(s): anagha.pradeep@research.iiit.ac.in;
Contributing authors: sriramk8@gmail.com; ketaki.shetye@research.iiit.ac.in; bassam.adnan@research.iiit.ac.in; radhika.mamidi@iiit.ac.in;

Abstract

Ambiguous words often require context-dependent sense disambiguation, making them a persistent challenge in several downstream Natural Language Processing (NLP) tasks such as Machine Translation (MT). Although modern MT models leverage contextual embeddings to capture semantic information, they often fail to correctly disambiguate the intended sense of ambiguous words from their surrounding context, resulting in semantically inaccurate translations. To address this challenge, we propose a sense-augmented parallel corpus for Sanskrit–English MT. We utilize our previously developed Sanskrit–English sense inventory, derived from the Monier–Williams dictionary and designed to support tasks such as word sense assignment and MT. Using this inventory, we augment the already existing Sanskrit–English parallel corpus (Itihāsa) with contextually appropriate word senses with the help of LLMs and qualitatively evaluate the word sense assignment capabilities of the LLMs. Furthermore, we investigate the impact of sense augmentation on MT by fine-tuning both the translation-focused Sarvam model and the general-purpose Llama 3.1 8B model. Our experimental results demonstrate that incorporating explicit sense information consistently improves translation quality by enhancing lexical disambiguation across models, with particularly notable benefits for low resource languages such as Sanskrit.

Keywords: Sense-Augmentation, Machine Translation, Word Sense Disambiguation (WSD), Corpus Construction

1 Introduction

Semantics is the branch of linguistics that studies meaning. It is one of the most multifaceted branches of linguistics, encompassing everything related to meaning, from its definition, relations, and constituents to the ways meanings combine and change (Riemer, 2010). Aside the philosophical inquiries of theoretical semantics, several aspects of semantics, including lexical, logical, compositional and cognitive, have been adapted for computational modelling in order to address real-world problems in NLP. The application of semantics in NLP has witnessed significant progress from formal semantics to computational semantics (more commonly known as distributional semantics), and within computational semantics from techniques such as one-hot encoding (bag-of-words), n-gram models, to static and dynamic word embeddings (Jurafsky & Martin, 2026).

While embeddings originated as a means of representing semantic relationships between words, they also capture multiple levels of linguistic information simultaneously. In addition to semantic properties, they have been shown to encode orthographic, morphological, lexical, and syntactic information, with the importance of each varying according to the model architecture, training objective, and downstream task (Adi, Kermany, Belinkov, Lavi, & Goldberg, 2016; Mikolov, Yih, & Zweig, 2013). Consequently, embedding models serve as general-purpose linguistic representations that can support a wide range of NLP tasks, ranging from low-level preprocessing tasks such as word segmentation and morphological analysis to higher-level applications such as Machine Translation (MT), question answering, and text classification. While this is true for both static and dynamic embeddings models, the dynamic embedding models capture these linguistic properties much more effectively because of contextual availability (Devlin, Chang, Lee, & Toutanova, 2019).

With respect to semantics, embedding models play an important role in representing the meanings of words. The process of determining the correct sense of a word in a given context is known as Word Sense Disambiguation (WSD) (Itankar & Raza, 2020). WSD has been an active sub-field in NLP, with applications in downstream tasks such as MT, Information Retrieval (IR), and Semantic Indexing. Since static embedding models represent each word with a single vector representation, they are unable to account for contextual variations in meaning, making separate WSD systems necessary for handling lexical ambiguity. On the other hand, the emergence of contextual embedding models has reduced the reliance on explicitly trained WSD systems, as these models are expected to encode word meaning directly from context. Nevertheless, the extent to which these dynamic models successfully distinguish between different senses of a word remains an open research question, particularly for low-resource languages. One way to assess their semantic capabilities is through their

performance on downstream tasks that depend heavily on accurate sense interpretation. In this work, we focus on MT as a representative downstream task for evaluating sense disambiguation in Sanskrit.

MT systems generally rely on parallel corpora for training. We investigate whether enriching the parallel corpus with explicit word sense information can improve the quality of MT model training. Specifically, this sense augmentation is performed through word sense assignment, where LLMs are used to identify and annotate the appropriate sense of each word based on its context.

Before presenting our framework and findings, we review existing data-driven research on Sanskrit WSD and MT in the next section on related work.

2 Related Work

2.1 Word Sense Disambiguation (WSD)

1. The first data-driven approach to WSD for Sanskrit was introduced by Hellwig (2017). His study focuses on assigning WordNet supersenses to rare and Vedic nouns extracted from the Digital Corpus of Sanskrit (DCS).[1] In this work, WSD is formulated as a sense classification problem, where the semantic category of a target noun is predicted using information from bilingual Sanskrit–English dictionary definitions along with its sentential context. The task is addressed at the word level using both Maximum Entropy (ME) and Recurrent Neural Network (RNN) classifiers.

   As a baseline, the system assigns the WordNet supersense associated with the headword of the first dictionary definition of a lemma. The ME model utilizes features derived from Sanskrit-English dictionary entries, including headwords, contextual words, and lexical information from neighboring Sanskrit lemmas within the sentence. To capture contextual relevance, surrounding lemmas are weighted according to their inverse distance from the target word. For the RNN model, English words are represented using pretrained GloVe embeddings, while out-of-vocabulary words are initialized randomly and updated during training. Sanskrit lemma representations are obtained through pretrained Word2Vec embeddings, which are incorporated into the classification framework.

   The experimental evaluation demonstrates that the RNN model consistently outperforms the ME classifier. For rare nouns, it achieves a micro-averaged F1 score of 77.68 and a macro-averaged F1 score of 55.71. For Vedic nouns, the model obtains a micro-averaged F1 score of 62.94 and a macro-averaged F1 score of 52.38.

   A major limitation of this work is that it performs WSD at the level of WordNet supersenses rather than fine-grained lexical senses, thereby providing only coarse semantic distinctions. Another limitation is the restricted scope of the dataset: the experiments are conducted on a relatively small subset of the DCS, focusing only on selected rare and Vedic nouns.
2. In 2025, Hellwig and Biagetti (2025) introduced a Sanskrit Sem Bank (SSB). The SSB combines lexicographic data from the DCS version of the Monier William's

[1] http://www.sanskrit-linguistics.org/dcs/

Sanskrit-English dictionary, along with synsets derived from the Princeton WordNet, and offers lexical semantic human annotations of more than 6,00,000 words in context across both Vedic and Classical Sanskrit texts. For portions of the corpus with limited annotation coverage, sense labels were assigned through simple WSD experiments. The experiments were conducted on a stratified sample of all polysemous words recorded in the SSB and evaluated using leave-one-out (LOO) cross-validation. A strong baseline was established using the Most Frequent Sense (MFS) heuristic. Several variants of this baseline were explored, including selecting a single MFS per subject category, with a fallback to the global MFS when only one sense was available. Similarly, text-specific and period-specific sense distributions, as well as combinations of text, subject, and subject-fallback strategies were also ventured. In addition, the study evaluated a Generalized Linear Model (GLM) for WSD. The best performance (86.7% F1) was achieved by a logistic GLM trained on dense contextual representations obtained by averaging the static embeddings of surrounding words, supplemented with dummy-encoded subject and text metadata.

3. Additional WSD experiments using the SSB was proposed by Hellwig, Sellmer, Nehrdich, Widmer, and Sennrich (2026). Instead of treating WSD as a traditional classification problem with one output label per word, the authors formulate it as a gloss-selection task i.e., given a Sanskrit sentence, a target lemma, and a set of candidate dictionary definitions derived from the Monier-Williams dictionary, the model must select the definition that best matches the word's meaning in context. To achieve this, the authors build a neural architecture inspired by gloss-reader models, where candidate definitions are encoded using SentenceTransformers, the target lemma is represented either through learned embeddings or a character-level transformer, and the sentence context is encoded with a pretrained Sanskrit character-level transformer. The model then learns to match a context representation with the correct definition embedding through similarity-based scoring. The paper's main contribution is demonstrating that lexicographic definitions themselves provide a powerful semantic inventory for Sanskrit WSD. Through a series of ablation experiments, the authors show that a trainable definition encoder, a character-level lemma encoder, and a context representation focused on the target-word region yield the best performance, achieving approximately 83.3% word-level accuracy on the SSB.

2.2 Machine Translation

Several efforts on MT for Sanskrit have been noted by Pradeep and Mamidi (2025). Among those, we only note the Machine Learning or Deep Learning based models, for the Sanskrit-English language pair here. A summary of these efforts is presented in Table 2.2.

1. Punia, Sharma, Pruthi, and Jain (2020) explore novel approaches previously unexplored in the translation of Sanskrit to English. Initially, the study establishes a baseline using the Transformer architecture with an encoder-decoder framework. Followed by this, advancements to the transformer model involve the integration

of Reinforcement Learning and Transfer Learning, resulting in a reported enhancement in BLEU scores attributed to the Transfer Learning mechanism. In addition to these methodological contributions, the study also provides an open dataset comprising both parallel and monolingual data. The parallel dataset, totaling 9,000 lines, is sourced from diverse sources, including 2,100 sentences from OPUS, Sanskrit translations of the Bible, ślokas from the Rāmāyaṇa, and additional sentences from the Bhagavad Gīta. The monolingual Sanskrit data consists of approximately 130,000 lines, utilizing the Romanized version of the Mahābhārata. For English monolingual data, the study extracts content from the Europarl dataset.

2. Koul and Manvi (2021) propose a neural MT system for translating Sanskrit to English, employing Recurrent Neural Networks (RNNs). The model uses vectors representing the source Sanskrit sentence, the target English sentence, and a bilingual dictionary. The RNN uses a computational graph to analyze words of the sentence in parallel, breaking them down into their roots and linguistic categories. The dependencies of the analyzed words concerning the context are determined and stored in the vector for subsequent processing. For simple sentences, words are translated using the bilingual dictionary, while for complex compounds, the learner module suggests translations. Partial sentences are generated, and context is used to adjust grammatical aspects like tense. Any additional extra-linguistic information if present is incorporated and finally the output is generated. The model employs Naive Bayes and Support Vector Machines in the classification algorithm to identify the most suitable target word. Notably, the study highlights that the use of a Support Vector Machine classifier in determining the appropriate target word, yielded higher accuracy compared to the Naive Bayes approach.
3. The work by Sitender and Bawa (2021) introduce a MT system designed for translating Sanskrit to English that employs a hybrid approach, combining both direct and rule-based translation techniques. The system addresses language divergence among Sanskrit and English languages and proposes a solution to handle it. The MT system incorporates two bilingual dictionaries (Sanskrit-English, Sanskrit-UNL), a tagged Sanskrit corpus, a Sanskrit analysis rule base, and an English Language Grammar Rule (ELGR) base. It comprises of six modules that include source language preprocessing, database, Sanskrit grammar demonstration, English parse tree generation, translation using the direct method, and output generation. For Sanskrit language processing, the system utilizes CFG in Chomsky Normal Form (CNF) and employs the Cocke–Younger–Kasami (CYK) parsing technique for processing input Sanskrit sentences. The article also introduces a novel algorithm that constructs a parse tree from the parsing table. The target language sentence is generated through the combined efforts of the ELGR base and bilingual dictionaries. Evaluation of the proposed system was conducted using the Natural Language Toolkit API in Python, resulting in a BLEU score of 0.7606, a fluency score of 3.63, and an adequacy score of 3.72.
4. Aralikatte, de Lhoneux, Kunchukuttan, and Søgaard (2021) introduce a large-scale translation dataset containing 93,000 pairs of Sanskrit śloka and their English translations extracted from two famous Indian epics, the Rāmāyana and the

Mahābhārata. The unabridged versions of these two texts along with their translations authored by M.N. Dutt have been taken acknowledging their large size and popularity. Further, this study attempts to highlight the complexity of the dataset by training and evaluating the performance of standard translation models. This includes one statistical model and five neural models. For statistical MT, Moses is used and for neural MT seq2seq Transformer with standard BERT architecture has been used. The evaluation is based on four parameters (i) character n-gram F-score (ii) token accuracy (iii) BLEU score and (iv) Translation Edit Ratio score. The study concludes that all the models performed poorly with low token accuracy and high translation edit ratio and underscores potential areas for improvement in the translation models.

5. Maheshwari et al. (2024) introduce Sāmayik, a parallel English-Sanskrit dataset comprising approximately 53,000 sentence pairs focused on contemporary Sanskrit prose. In contrast to existing Sanskrit corpora, which predominantly contain classical literature and poetry, Sāmayik is curated from diverse contemporary sources, including language learning materials, pedagogical texts, and online tutorials. The authors highlight the scarcity of digitized contemporary Sanskrit resources and demonstrate that MT models trained on Sāmayik significantly outperform models trained on poetry-centric datasets when evaluated on out-of-domain contemporary Sanskrit corpora. The work employs multilingual sequence-to-sequence models such as ByT5, mBART, dicBART, and IndicTrans for bidirectional English-Sanskrit MT. Among these, IndicTrans was pre-trained on Sanskrit data, whereas ByT5, mBART, and dicBART had not been exposed to Sanskrit during pre-training and were subsequently adapted for the translation task. The authors also release these as benchmark models for bidirectional English-Sanskrit neural MT.

Besides these standalone models, multilingual translation models like Google Translate,[2] IndicTrans (multiple versions),[3] and Sarvam[4] also include Sanskrit among their supported languages. While Google Translate and IndicTrans have been used and evaluated for Sanskrit-English translation by Maheshwari et al. (2024) and Pradeep, Mamidi, and Satuluri (2024), Sarvam, which is based on the Gemma architecture has not yet been assessed for this language pair.

3 Motivation

Despite advances in Sanskrit WSD, the impact of these improvements on downstream applications remains largely unexplored. In particular, MT, which critically depends on accurate sense assignment, has not been systematically evaluated from a semantic disambiguation perspective. Existing Sanskrit–English MT systems, including both standalone translation models and multilingual systems, have primarily been assessed using surface-level metrics such as BLEU and chrF (refer Table 2.2). While these metrics measure lexical and n-gram overlap, they do not adequately capture whether

[2] https://translate.google.co.in/?sl=sa&tl=en&op=translate
[3] https://github.com/AI4Bharat/indicTrans https://huggingface.co/collections/ai4bharat/indictrans2
[4] https://huggingface.co/sarvamai/sarvam-translate

Work	Models	Data	Metrics
Punia et al. (2020)	Seq2Seq with Attention, Transfer Learning, and Reinforcement Learning	Parallel Sanskrit–English corpus with additional monolingual Sanskrit data. Transfer learning from high-resource language pairs is also used.	BLEU
Koul and Manvi (2021)	Hybrid system combining RNN-based NMT, SVM, Naïve Bayes, bilingual dictionary, and rule-based preprocessing	Tagged Sanskrit corpus, bilingual dictionary, and manually prepared examples. No publicly released parallel corpus; corpus size not reported.	Accuracy, Translation speed
Sitender and Bawa (2021)	Hybrid (Direct + Rule-Based) MT using CFG (CNF), CYK parser, handcrafted grammatical rules, Sanskrit–English and Sanskrit–UNL dictionaries, ELGR, and Elasticsearch	Tagged Sanskrit corpus, handcrafted grammatical rules, and bilingual dictionaries. No neural training corpus or train/test split reported.	BLEU, Human Fluency, Human Adequacy
Aralikatte et al. (2021)	Transformer-based NMT models: Moses, B2B-Tiny, B2B-Mini, B2B-Small, B2B-Medium, and B2B-Base	*Itihasa*: approximately 93,000 Sanskrit-English śloka pairs extracted from the *Rāmāyaṇa* and *Mahābhārata*.	BLEU, chrF
Maheshwari et al. (2024)	Fine-tuning multilingual pretrained translation models: ByT5, mBART, IndicBART, and IndicTrans	*Sāmayik*: approximately 53,000 Sanskrit–English parallel sentences consisting mainly of contemporary Sanskrit prose from diverse domains.	BLEU, chrF

Table 1 Summary of Sanskrit–English Machine Translation Systems

a system has correctly interpreted and translated the intended sense of a polysemous word.

This gap became evident in our earlier qualitative analysis of Sanskrit translation systems, specifically Google Translate Pradeep et al. (2024), where we observed that modifying the surrounding context often failed to induce the expected changes in translation. It was observed how polysemous words continued to be translated with the same sense irrespective of contextual cues, suggesting that the models were either unable to perform context-sensitive disambiguation or had not adequately learned

alternative senses during training.[5] These observations raised a more fundamental question: do current translation models possess robust representations of sense distribution, and what kinds of training data or evaluation methodologies are required to improve their sense assignment capabilities?

Addressing this question requires training data that extend beyond just parallel corpora and evaluation methods that go beyond the surface-level automatic metrics. Motivated by these gaps, we investigate an alternative method for data enhancement as well as other evaluation metrics that can provide greater insight into the semantic adequacy of translated outputs.

Furthermore, while Sarvam provides an LLM-based architecture for translation, its performance has not been systematically evaluated for Sanskrit. Therefore along with Sarvam, we include other LLMs in our study to examine how they perform on Sanskrit translation in comparison with dedicated MT systems.

This work proposes a framework for training Sanskrit–English MT systems with a particular focus on lexical ambiguity and word-sense assignment. Using this framework, we evaluate the performance of Google Translate, IndicTrans, Sarvam, and Llama 3.1 8B in three settings: **(1) zero-shot, (2) fine-tuning on parallel corpora, and (3) fine-tuning using the proposed framework**. This enables us to assess the extent to which targeted training for lexical ambiguity improves sense-aware translation across different model architectures. Figure 1 provides an overview of the proposed methodology, including dataset construction, sense augmentation, model training, and evaluation. It serves as a roadmap for the remainder of the paper.

The rest of this paper is organized as follows. Section 4 describes the creation of the dataset and the proposed framework. Section 5 presents the experimental setup and results. Section 6 provides an error analysis of the translation outputs. Section 7 discusses the limitations of the proposed approach. Finally, Section 8 concludes the paper and outlines directions for future work.

4 Data Creation

A fundamental requirement for any MT system is the availability of parallel corpora for training and evaluation. Very limited parallel data does exist for the Sanskrit–English language pair. In this work, we use the Itihāsa corpus by Aralikatte, Gantayat, Panwar, Sankaran, and Mani (2018), which contains 93,030 pairs of Sanskrit ślokas and their English translations drawn from the two major epics, the Rāmāyaṇa and the Mahābhārata.

A key issue with this corpus was a mismatch in the alignment between source and target sentences. We address this by performing manual verification and cleaning. Approximately 75–100 sentence pairs were manually realigned, and 30 pairs were removed due to irrecoverable misalignment, resulting in a final dataset of 93,000 aligned sentence pairs. We call this our **base dataset**. This dataset forms the foundation for our proposed framework, which is designed to explicitly train and evaluate sense assignment in Sanskrit–English MT.

[5] Most of these issues continue to be observed even in the recent releases of Google Translate.

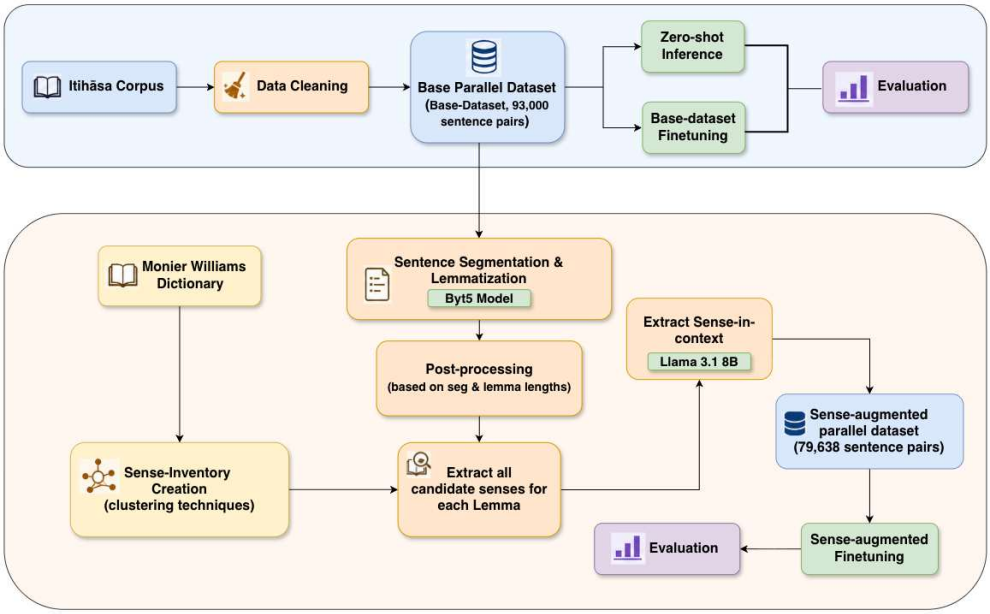


Fig. 1 Overview of the proposed framework. The upper pipeline represents the base-dataset creation and translation setup, while the lower pipeline illustrates the proposed sense-augmentation framework used to construct a sense-augmented parallel corpus for model fine-tuning.

4.1 Proposed Framework: Resource Construction

This work is motivated by the hypothesis that explicitly incorporating sense distinctions in the training data, similar to a sense inventory, can improve a model's ability to perform context-sensitive translation. To this end, we set out to create a sense inventory for Sanskrit. At the time of initiating this work, no SemBank-style resource was available for Sanskrit and existing WordNets were also incomplete. The only lexical resource with sense-level annotations was the DCS corpus. However, even that was partially complete. Consequently, we explored available lexical resources that could serve as a foundation for constructing a sense inventory for Sanskrit. And we chose the Monier-Wiiliam's (MW) Sanskrit-English dictionary as our resource for inventory creation. Rather than relying on the original XML representation of the dictionary, we make use of an aligned and structured version developed by Patel and Kulkarni (2024).

Our approach was to extract all possible senses for each word of each sentence in the Itihāsa corpus from the MW dictionary and use LLMs to select the sense-in-context, based on the translation and the full set of derived dictionary senses. In order to do this, we first extract lemmas from the Sanskrit verses in the dataset. We perform lemma extraction along with segmentation using the ByT5-Sanskrit model by Nehrdich, Hellwig, and Keutzer (2024). We post-process the dataset by removing sentence pairs in which there is a mismatch between segmentation length and lemmatization length. We also discard sentence pairs where the segmentation length

is shorter than the sentence length. Only those sentence pairs with equal segmentation and sentence lengths, or with segmentation length greater than the sentence length, are retained. This filtering is performed after segmentation and lemmatization to exclude cases where not all tokens are associated with corresponding lemmas, ensuring complete lemma coverage.[6] This results in a final post-processed dataset of **79,638** sentence pairs (refer Table 2 for complete statistics).

Category	**Number of Sentence Pairs**
Segmentation–lemma length mismatch	8,800
Segmentation length $<$ sentence length	4,562
Segmentation length $=$ sentence length	441
Segmentation length $>$ sentence length	79,197
Total dataset size	93,000
Final retained dataset	79,638

Table 2 Statistics of dataset filtering during post-processing.

Using the extracted lemmas, we retrieve all corresponding senses from the Monier-Williams (MW) dictionary by mapping each lemma in the corpus to its dictionary entry. In cases where no corresponding entry is found in the MW dictionary, we obtain the senses from the DCS dictionary.

After extracting the MW dictionary senses for each lemma, we observed that the dictionary is designed for human readers, making its structure ambiguous for computational use. A key challenge is handling entries with multiple senses or nested sub-entries, especially for verbs and indeclinables. This raises important questions: should such entries be split into separate sense records, and if so, based on what criteria? Or should they remain unified, with mechanisms to identify and reference individual senses within a single entry?

To better understand the dictionary's organization, we examined its preface, which states that entries are structured according to both grammatical properties and semantic distinctions. Homonymous headwords are listed separately, while semantically distinct (polysemous) senses are separated by semicolons and closely related senses by commas. In the aligned version used in this study, semicolon-separated senses are represented as separate entries, whereas comma-separated senses remain grouped.

However, our analysis revealed that these principles are not applied consistently across parts of speech, particularly for nouns, verbs, and indeclinables. We also found that many entries include grammatical conditions (e.g., tense, aspect, and mood) as sub-keys that introduce additional sense distinctions. Moreover, several entries labeled as polysemous did not appear to represent genuinely distinct senses, while many grouped synonyms were not truly synonymous.

To address these inconsistencies, we systematically analyzed the dictionary entries using cosine similarity over sense embeddings generated by five encoder models. We

[6] While our intention is to extract the lemmas alone, we perform segmentation as well primarily to assist the post-processing step.

then applied four clustering algorithms to compare and evaluate sense groupings. Based on these findings, we reorganized the dictionary structure by transforming the human-oriented resource into a machine-usable sense inventory such that it better supports downstream NLP applications. We refer to this resource as the **sense-inventory-base**. Additional details on its construction are provided in our companion paper Pradeep, Krishnan, and Mamidi (2026).

As the next step towards enhancing the sense-inventory-base, we used LLMs to extract the sense-in-context for each lemma across the 79,638 sentences of the Itihāsa corpus. For this purpose, we use Llama 3.1 8B Instruct.[7] For each sentence, we provide the model with the sentence, its English translation, all lemmas occurring in the sentence, and the candidate senses for each lemma (extracted from the sense inventory). The model is then prompted to select the sense-in-context for each lemma, using the English translation and the candidate sense list as disambiguating context. This is done in a completely zero-shot setting, without any task-specific fine-tuning or in-context examples.[8]

To validate the approach, we manually evaluated the model's predictions on a sample of 30 sentences (476 lemmas) and found the model performed well, achieving an accuracy of 93.07%. With sense-in-context labels assigned across the full corpus, the resulting dataset is enriched with explicit, fine-grained sense annotations for every lemma occurrence. We refer to this as the **sense-augmented Itihāsa corpus**, and we show the application of this resource in our experiments.

4.1.1 Statistics of the Sense-Augmented Itihāsa Corpus

The sense-augmented Itihāsa corpus contains 12,83,494 lemma occurrences spanning 21,393 unique lemmas. Across these lemmas, the underlying sense inventory contains 1,51,790 distinct senses. The number of senses associated with a lemma varies considerably, ranging from a single sense to 138 senses for the most polysemous lemma. Overall, a lemma is associated with an average of 7.1 senses, while the median senses per lemma is 4. In terms of lemma-level sense coverage, the average coverage is 0.486, indicating that, on average, 48.6% of the senses associated with a lemma are attested in the corpus. At the corpus level, however, the distribution is more conservative. Of the 1,51,790 distinct senses in the underlying inventory, only 58,281 are attested in the corpus, whereas 93,509 remain unattested. Thus, 38.4% of the unique senses are covered by the corpus, while 61.6% are unused. The main corpus-level statistics are summarized in Table 3.

The distribution of the number of associated senses per lemma is shown in Table 4. A total of 3,480 lemmas (16.27%) are associated with a single sense, while 3,999 (18.69%) have two senses and 2,480 (11.59%) have three senses. Thus, 46.55% of the lemmas have three or fewer associated senses. At the other end of the distribution, 4,214 lemmas (19.70%) are associated with more than ten senses. However, a visual inspection (in Figure 2) indicates a skewed distribution in which a relatively small number of lemmas account for a substantial number of sense distinctions.

[7] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
[8] The prompt template for sense extraction is given in Appendix A.

Statistic	Value
Lemma occurrences	12,83,494
Unique lemmas	21,393
Distinct senses	1,51,790
Average senses per lemma	7.1
Median senses per lemma	4
Minimum senses per lemma	1
Maximum senses per lemma	138
Lemma-level sense coverage	0.486
Lemma-level sense unused	0.514
Unique used senses	58281
Unique unused senses	93509
Corpus-level unique sense coverage	0.384
Corpus-level unique sense unused	0.616

Table 3 Basic statistics of the sense-augmented Itihāsa corpus.

Lexical categories such as prātipadika (nominal stems), cpd-prātipadika (compounds), dhātu (verbal roots) and so on were assigned for each of the senses in the MW dictionary semi-automatically (refer Pradeep et al. (2026) for more information). The same categorization was applied to the sense-inventory-base and subsequently carried over to the sense-augmented corpus. This resulted in diverse lexical composition of the corpus across these categories. Compound nominal lemmas (cpd-prātipadika) constitute the largest category, accounting for 43.08% of all unique lemmas, followed by prātipadika (30.53%). The remaining categories comprise a relatively small proportion of the dataset. Table 5 summarizes the number of lemmas and associated senses for each category, along with the average number of senses per lemma.

The distribution of senses also varies across the different lexical categories. For example, dhātu and upasarga contain relatively few lemmas but have higher average numbers of senses per lemma than the larger categories. This variation highlights the differences in semantic complexity across lexical categories.

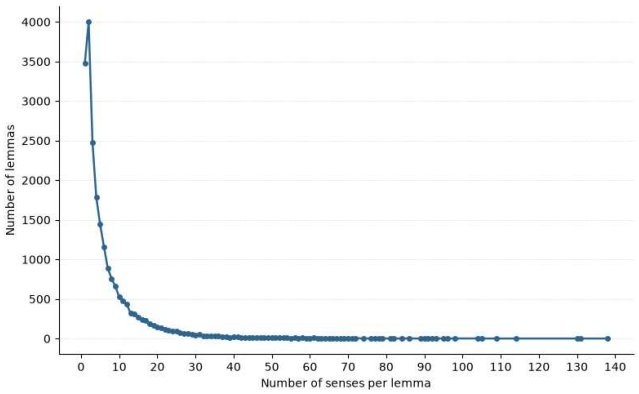


Fig. 2 Distribution of lemmas by number of associated senses.

Senses per lemma	No. of lemmas	Percentage
1	3,480	16.27
2	3,999	18.69
3	2,480	11.59
4	1,784	8.34
5	1,443	6.75
6	1,162	5.43
7	891	4.16
8	757	3.54
9	660	3.09
10	523	2.44
> 10	4,214	19.70

Table 4 Distribution of lemmas by number of associated senses.

The table also indicates the senses obtained from DCS. It is important to note that DCS does not represent a distinct lexical category; rather, it denotes cases where the required senses were unavailable in the aligned version of MW and were therefore obtained from DCS.

Category	Lemmas	Senses	Avg. senses/lemma
cpd-prātipadika	9,216	63,439	6.88
prātipadika	6,531	64,508	9.88
dhātu	485	10,608	21.87
cvi-cpd	78	866	11.10
avyaya	57	734	12.88
nāma-dhātu	51	317	6.22
cpd-avyaya	41	182	4.44
upasarga	17	441	25.94
DCS	5,156	10,732	2.08

Table 5 Distribution of lemmas and senses across lexical categories.

Taken together, these statistics provide an overview of the size, sense distribution, and lexical composition of the sense-augmented Itihāsa corpus.

5 Experiments

As noted in the introduction, the semantic capabilities of embedding models, which are strongly influenced by the training data, can be evaluated by examining their performance on downstream tasks such as MT that rely heavily on accurate sense interpretation. Accordingly, our experiments evaluate various Transformer-based MT models trained under three different stages with varying data settings.

5.1 Models

We evaluate four transformer-based MT systems: Google Translate[9] (a proprietary, closed-source NMT system), IndicTrans2[10] (a multilingual transformer model for Indic languages), Sarvam Translate[11] (multilingual model for Indian languages built on Gemma), and Llama 3.1 8B Instruct (a general-purpose instruction-tuned LLM).

5.2 Data Splits

We split our base dataset (the cleaned parallel Sanskrit–English corpus described in Section 4) into a training set of 79,050 sentence pairs and a test set of 13,949 sentence pairs.

Similarly, we split our sense-augmented Itihasa dataset of 79,638 samples into a training set of 67,624 samples pairs and a test set of 12,014 samples.

5.3 Experimental Stages

We conduct experiments in three stages:

1. **Zero-shot inference:** We first evaluate all four models directly on the test set of the base dataset without any fine-tuning, to establish baseline performance.
2. **Fine-tuning on the base dataset:** We fine-tune the three open models — Llama, Sarvam, and IndicTrans2 — on the training split of the base dataset. Google Translate is excluded from this stage as it is not an open source model.
3. **Fine-tuning on the sense-augmented Itihāsa corpus.** We also fine-tune the models on the training set of our sense-augmented Itihāsa corpus, where each training instance consists of the sentence, its lemmas, and the sense-in-context for each lemma, and the model is prompted to produce the translation. Since this stage relies on instruction/prompt-based fine-tuning, IndicTrans2, which is not an instruction-tuned model, is excluded; we report results only for Llama and Sarvam.[12]

6 Results

We evaluate each stage of the experiments using five (5) MT evaluation metrics: BLEU, chrF, COMET, MEE4, and chrF-S.

BLEU Papineni, Roukos, Ward, and Zhu (2002) and chrF Popović (2015) are reference-based metrics relying on n-gram overlap, while COMET Rei, Stewart, Farinha, and Lavie (2020), is a fully supervised metric, fine-tuned on large-scale human judgment data, to directly predict a human-like quality score, without relying on explicit n-gram matching. MEE4 Mukherjee and Shrivastava (2023) is an unsupervised, reference-based metric that leverages embeddings to quantify lexical, syntactic, semantic, morphological, and contextual similarity between hypothesis and reference. chrF-S, introduced in Mukherjee and Shrivastava (2024), is a distinct metric from chrF,

[9] https://docs.cloud.google.com/translate
[10] https://huggingface.co/ai4bharat/indictrans2-indic-en-1B
[11] https://huggingface.co/sarvamai/sarvam-translate
[12] The prompt template for instruction fine-tuning is given in Appendix A.

combining character/word n-gram matching with sentence embedding-based semantic similarity to capture both syntactic accuracy and sentence-level meaning.

The results for each experimental stage are reported using these five evaluation metrics. BLEU and chrF are reported on their standard 0-100 scales, whereas COMET, MEE4, and chrF-S are reported in their original scoring scales. For all metrics, higher scores indicate better translation quality.

1. **Zero-shot Inference:** Table 1 reports zero-shot performance of all four models on the test set. Google Translate achieves the strongest performance across most metrics, followed by Sarvam, while Llama and IndicTrans2 lag behind, particularly on BLEU and chrF.

Model	BLEU	chrF	COMET	MEE4	chrF-S
Google	**6.5010**	**32.2595**	**0.6716**	**0.5743**	**0.4718**
IndicTrans2	2.5822	19.0072	0.5220	0.3560	0.3140
Sarvam	5.7206	29.2328	0.6485	0.5247	0.2239
Llama 3.1	2.5779	19.7077	0.4962	0.3498	0.3152

Table 6 Performance comparison of baseline translation models. Higher values indicate better performance for all metrics.

2. **Fine-tuning on the Base Dataset:** Table 2 shows results after fine-tuning Llama, Sarvam, and IndicTrans2 on the training split of the base dataset. All three models show substantial improvement over their zero-shot counterparts, with Sarvam and Llama achieving comparable BLEU scores and consistent gains across chrF and COMET as well.

Model	BLEU	chrF	COMET	MEE4	chrF-S
IndicTrans2	10.3280	35.6571	0.6683	0.5895	0.5228
Sarvam	11.8588	37.6901	0.6937	0.6152	**0.5459**
Llama 3.1	**11.9588**	**38.1850**	**0.6982**	**0.6188**	0.2877

Table 7 Performance comparison of translation models on the finetuned base dataset. Higher values indicate better performance for all metrics.

3. **Fine-tuning on the Sense-Augmented Itihāsa Corpus:** Table 3 presents results after further fine-tuning Llama and Sarvam on the sense-augmented Itihāsa corpus, using sentence, lemma, and sense-in-context as input. Both models show additional gains over the base fine-tuned setting across all metrics, with the most notable improvement in chrF-S, particularly for Llama, which improves from 0.29 to 0.57, nearly doubling its score.

Across all three stages, a clear and consistent trend emerges: fine-tuning on the base dataset substantially improves performance over zero-shot inference across all trainable models, with the exception of the chrF-S score for Llama. Fine-tuning

Model	BLEU	chrF	COMET	MEE4	chrF-S
Sarvam	**13.6889**	**41.2287**	0.7098	0.6412	0.5734
Llama 3.1	13.5535	40.5354	**0.7120**	**0.6413**	**0.5752**

Table 8 Performance comparison of translation models on the sense-augmented Itihāsa dataset. Higher values indicate better performance for all metrics.

on the sense-augmented corpus yields further gains across nearly every evaluation metric. These improvements are particularly pronounced for COMET and chrF-S, metrics designed to capture semantic correctness, indicating that the incorporation of explicit sense information enables models to make more contextually appropriate lexical choices during translation. Overall, these findings support our original hypothesis that sense-augmented MT is an effective approach for improving semantic accuracy and overall translation quality.

7 Discussion

Having presented the experimental results, this section discusses the key findings in the context of the proposed sense-aware MT pipeline. As the challenges associated with the base parallel corpus (base dataset) have already been discussed in Section 4, this discussion focuses on the resource construction process, including the development of the sense-inventory-base and the creation of the sense-augmented corpora. It then examines the experimental results and analyzes the effectiveness of the overall translation pipeline.

7.1 Resource Construction

1. **Sense-Inventory-Base:**

   - The primary challenge in constructing the sense-inventory-base stemmed from the nature of the MW dictionary, which is computationally noisy for sense-based tasks. Its entries contain definitions, citations, grammatical markers, and highly variable sense lengths, making the extraction of individual senses challenging. While this additional information can support disambiguation, isolating clean and consistent sense definitions remains difficult. Consequently, the models and algorithms employed to restructure the MW dictionary and construct the sense inventory introduced their own advantages and limitations, which are discussed in detail in our previous work (Pradeep et al., 2026).
   - According to the results presented in our previous work (Pradeep et al., 2026), the LaBSE embedding model combined with the Louvain clustering algorithm achieved the best performance across the evaluated approaches. Therefore, we selected this model-algorithm combination for constructing the sense-inventory-base. While the results demonstrate that this approach outperformed the alternatives, we identified one significant limitation; this model-algorithm combination generated a large number of clusters, resulting in an extremely fine-grained

sense inventory. While this granularity allows for more detailed sense distinctions, it may also introduce additional ambiguity for downstream tasks by making sense selection more difficult. This limitation became evident during sense-in-context extraction using Llama 3.1 8B, where the model was required to distinguish among a large number of closely related senses.

2. Sense-Augmented-Itihāsa-Corpora

   - We used ByT5-Sanskrit, the current state-of-the-art model for Sanskrit segmentation and lemmatization, to predict the lemmas for all sentences in the base dataset. In addition to ByT5-Sanskrit, we explored alternative segmentation systems, specifically the Sanskrit Heritage Platform (SH) (Gérard, 2003), which employs a probabilistic ranking approach to select the most likely segmentation and morphological analysis from all possible solutions generated by a Finite State Transducer (FST) (Krishnan, Kulkarni, & Huet, 2025). A major limitation of SH is its dependence on a predefined lexicon, making it susceptible to out-of-vocabulary (OOV) words. Consequently, among the 93,000 sentences processed, approximately 24,000 contained at least one word that could not be recognized by the engine. In contrast, using ByT5-Sanskrit required pruning only about 14,000 sentences.
     Comparisons between ByT5-Sanskrit and SH have received limited attention in the literature. One such study by Venkatesh et al. (2026), evaluates the two systems on a manually annotated corpus of 3,669 sentences[13] available through Saṃsādhanī's START platform.[14] In that evaluation, SH achieved a slightly higher F1-score than ByT5-Sanskrit (82.70% compared to 80.33%). Despite this 2.37 percentage-point increase in F1-score, we selected ByT5-Sanskrit because of its higher recall, which was more beneficial for the segmentation and lemmatization requirements of our base dataset. We also observed systematic differences between the lemmas produced by ByT5-Sanskrit and those listed in the MW dictionary. ByT5-Sanskrit is primarily trained on the DCS and therefore follows the DCS annotation conventions for lemmatization. Unlike the format followed in MW, DCS may assign either a base lemma or a derived lemma depending on the context. As a result, some derived lemmas generated by ByT5-Sanskrit are not present in the MW dictionary. Furthermore, the two resources differ in their treatment of pronouns. While MW adopts the traditional grammatical stems *asmad* and *yuṣmad* as lemmas, DCS instead uses the forms *mad* and *tvad*. These differences require additional normalization when aligning ByT5-Sanskrit-generated lemmas with MW dictionary entries.
   - The long-tailed distribution of senses across lemmas, as observed in the sense-augmented corpus, reflects the fine-grained nature of the constructed sense inventory and presents a practical challenge for sense-in-context extraction. Therefore, to evaluate the utility of the sense-inventory-base, we randomly took a sample of 30 sentences from our corpus and prompted Llama 3.1 8B to perform

[13] The corpus includes simple prose sentences, classical literary texts such as the Bhagavad Gītā and Saṅkṣepa Rāmāyaṇa, and the domain-specific text Aṣṭāṅgahṛdayam.
[14] https://start.samsaadhanii.in/

sense-in-context extraction using two lexical resources: the original MW dictionary and the constructed sense-inventory-base. The extracted senses were then manually validated. As shown in Table 2, the MW dictionary achieved a slightly higher extraction accuracy, correctly handling 451 out of 472 lemmas (95.55%), compared to 443 out of 476 lemmas (93.07%) for the sense-inventory-base.

Model	Resource	Correct	Missed	Total Lemmas
Llama 3.1 8B	MW Dictionary	455 (95.17%)	23 (4.83%)	476
Llama 3.1 8B	Sense Inventory	443 (93.07%)	33 (6.93%)	476

Table 9 Sense extraction performance of Llama 3.1 8B using different lexical resources.

Although the MW dictionary yielded an approximately 2% higher extraction accuracy, it proved considerably less practical for large-scale processing. Processing the same set of 30 sentences required nearly twice the inference time compared to the sense-inventory-base. Furthermore, MW entries contain rich but extensive lexical information (notes) as mentioned earlier, which substantially increases the prompt length and context requirements for LLM-based extraction. In contrast, the sense-inventory-base provides a compact gloss-oriented representation that is more efficient for downstream tasks.

Based on these observations, we adopted the sense-inventory-base for all subsequent experiments. Although the resource remains highly fine-grained and could benefit from future refinement into a coarser-grained inventory, it offers a practical trade-off between extraction accuracy and computational efficiency, making it better suited for downstream sense-aware MT.

We also analyzed the distribution of extraction errors across the different lexical categories. For the original MW dictionary, most errors involved prātipadikas (12 instances), followed by compound prātipadikas (5), dhātus (3), and avyayas (1). Using the sense-inventory-base, errors increased across all major categories, with prātipadikas accounting for 15 instances, followed by dhātus (9), compound prātipadikas (7), avyayas (2), and upasargas (2). The largest increase was observed for dhātus, suggesting that verb sense selection is particularly sensitive to the finer-grained distinctions introduced by the automatically constructed sense inventory. In contrast, the relatively small increase in errors for nominal categories indicates that the sense inventory remains reasonably effective for noun sense extraction despite its higher level of granularity.

7.2 Model Performance

Across all experiments, model performance improved progressively from zero-shot inference, to fine-tuning on the base dataset, and finally to fine-tuning on the sense-augmented corpus. This trend was consistent across almost all evaluation metrics. One notable exception was the chrF-S score for Llama 3.1 8B, which showed a slight decline after fine-tuning on the base dataset relative to the zero-shot model.

We did some qualitative analysis that suggests that this decrease is primarily due to degenerate generation, where the model enters repetitive loops and repeatedly generates the same phrase or sentence. For instance,

1. **Example 1:** And sighing, and racked with grief, and her heart overwhelmed with ire, Mithila's daughter, afflicted, shed tears, and bewailed, saying, "O Rāma! O Lakṣmaṇa! O Kausalyā! O Sumitrā! O my mother-in-law!
2. **Example 2:** Having said this, that repressor of foes, placing his bow in the middle, set up a loud twang, filling all the points of the compass with the sound. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were greatly terrified. And the Tāṭakas, hearing that sound, were

As illustrated in the examples above, the model correctly begins the translation but subsequently repeats portions of the output (e.g., repeated invocations such as "O my mother-in-law!" or entire clauses such as "And the Tāṭakās, hearing that sound, were greatly terrified.") until the generation terminates. Although the initial portion of the translation is largely faithful to the source, these repetitions substantially increase the edit distance from the reference translation, thereby lowering the chrF-S score. We observed this behaviour in 347 generated outputs. Such repetition is a well-known characteristic of autoregressive language models during generation. However, this repetition was substantially reduced after fine-tuning on the sense-augmented corpus, leading to improvements in chrF-S primarily.

Nevertheless, the observed performance gains cannot be attributed solely to the reduction in repetitive generation. Improvements are evident across the dataset, with higher translation quality reflected in the scores for individual sentences as well as in the overall average across all evaluation metrics. To illustrate these improvements, we present examples that highlight the impact of accurate sense assignment on translation quality.

Example 1:

- **Source:** tatastu kṣatriyāḥ kecij jāmadagnyabhayārditāḥ |viviśur giridurgāṇi mṛgāḥ siṃhārditā iva ||
- **Reference:** Then, some of the Kshatriyas, stricken with the terror of Jamadagni's son, entered mountain fastnesses, like deer afflicted by the lion.
- **Baseline (Zero-shot Inference):** kecijjāma
- **Fine-tuning on the Base Dataset:** Then some Kshatriyas, terrified by the fear of Jamadagni's son, entered the forest, like deer entering a forest infested by lions.
- **Fine-tuning on the Sense-Augmented Dataset:** Then some Kshatriyas, afflicted with the terror of Jamadagni's son, entered the mountain strongholds like deer afflicted with lions.

As shown in the example, the baseline model produced a highly random output consisting of only a single word. After fine-tuning on the parallel corpus, the model was able to generate a valid and coherent sentence. However, the term "*giridurgāṇi*" was translated as "forest" in the baseline dataset fine-tuned model, which is incorrect in this context. The reference translation renders it as "mountain fastness", while the sense-augmented fine-tuned model translates it as "mountain strongholds", a semantically much closer rendering.

The improvement is also reflected in the semantic evaluation metrics: the COMET score increases from 0.2856 for the zero-shot inference to 0.7946 after parallel corpus fine-tuning and further to 0.8198 with sense-augmented fine-tuning. Similarly, the chrF-S score rises from 0.0499 to 0.7095, and finally to 0.8017. The following examples also show similar improvements.

Example 2:

- **Source:** skandhāvāreṇa mahatā kurukṣetraṃ jagāma ha |
- **Reference:** And with a large army he went to Kurukshetra.
- **Zero-shot Inference:** He
- **Fine-tuning on the Base Dataset:** He then went to Kurukshetra, being covered with a large quantity of cloth.
- **Fine-tuning on the Sense-Augmented Dataset:** With a mighty army he went to Kurukshetra.

In this example, the word "*skandhāvāreṇa*" has been incorrectly translated as "cloth" by the base dataset fine-tuned model. This translation is entirely inconsistent with the context, where the correct meaning is "army". Both the reference translation and the sense-augmented fine-tuned model accurately translate it as army.

Here, the COMET score improves from 0.3456 for the zero-shot inference to 0.5730 after parallel corpus fine-tuning, and further increases significantly to 0.9195 with sense-augmented fine-tuning. Similarly, the chrF-S score improves from 0.0467 to 0.1988, and finally to 0.7857 across the three stages.

Example 3:

- **Source:** mā viṣādaṃ gamas tasmān naitat tvayy upapadyate |vinaṣṭe tvayi cāsmākaṃ pakṣo hīyeta kaurava ||
- **Reference:** Do not give way to despondency. O descendant of Kuru, this conduct does not suit you. If you die, our party will become weak.
- **Zero-shot Inference:** gamastasmān
- **Fine-tuning on the Base Dataset:** Do not grieve, O son of Kuru, for this cannot happen to you. If you are destroyed, our cause will be lost.
- **Fine-tuning on the Sense-Augmented Dataset:** Do not give way to grief. It is not proper for you to do so. O descendant of Kuru, if you perish, our party will be weakened.

In this example, the word "*pakṣa*" is translated as "cause" by the base dataset fine-tuned model, which is incorrect in the given context. The reference translation and the sense-augmented fine-tuned model both correctly translate it as "party". Also, the word "*Kaurava*" has been translated as the "son of Kuru" in the base-dataset fine-tuning, while it has been translated correctly as "descendant of Kuru" in the sense-augmented fine-tuning.

For this example, the COMET score increases from 0.2819 for the baseline model to 0.7347 after base dataset fine-tuning, and further improves to 0.8312 with sense-augmented fine-tuning. Similarly, the chrF-S score improves from -0.0095 to 0.1894, and finally to 0.7331 across the three stages.

The results across the three examples, highlight the effectiveness of sense augmentation in resolving lexical ambiguity and generating translations that are more contextually appropriate and semantically aligned.

With respect to Sarvam, a translation-oriented model, consistent improvements were observed across the different stages of training, with performance increasing progressively from zero-shot inference to parallel corpus fine-tuning and finally to sense-augmented fine-tuning. However, the improvement achieved through sense-augmented fine-tuning over parallel corpus fine-tuning was relatively modest (approximately 3%), suggesting that Sarvam was already capable of producing high-quality translations after conventional fine-tuning on the parallel corpus. Sense augmentation further enhanced its performance by providing additional semantic refinements.

On the other hand, Llama 3.1 8B, being a general-purpose LLM rather than a translation-specific model, benefited more substantially from sense-augmented fine-tuning. The additional semantic supervision enabled Llama 3.1 8B to significantly improve its translation capabilities and achieve performance comparable to Sarvam. These results demonstrate that sense augmentation serves as an effective complementary training strategy, not only for further improving dedicated translation models but also for adapting general-purpose LLMs to specialized translation tasks.

All models were trained for a total of eight epochs, and the best-performing checkpoint was selected based on validation performance. As shown in Table 7.2, the optimal checkpoint was consistently reached before the final epoch, indicating that convergence occurred early during training. This observation suggests that extending training beyond these epochs did not yield further improvements and may have led to performance saturation. We can also observe how both models exhibited distinct convergence

behaviors during fine-tuning. Llama 3.1 achieved its best performance after five to six epochs, indicating that it benefited from prolonged optimization before reaching peak validation performance. In contrast, Sarvam consistently attained its best checkpoint within the first two to three epochs across both experimental settings. This suggests that Sarvam converged more rapidly on the task, likely owing to better alignment between its pretrained representations and the target domain.

Model	Total Epochs	Best Epoch (Step)
Fine-tuning 1 – Llama 3.1 8B	8	6 (53,364)
Fine-tuning 2 – Llama 3.1 8B	8	5 (38,040)
Fine-tuning 1 – Sarvam	8	2 (71,146)
Fine-tuning 2 – Sarvam	8	3 (91,293)

Table 10 Training epochs and best-performing checkpoints for Llama 3.1 and Sarvam.

It is also important to note that although the Itihāsa corpus serves as the primary data source for our experiments, unlike the original benchmark setup, we use our own data split to evaluate the translation models and their fine-tuned variants. Consequently, the reported scores are intended to provide a comparative analysis of these models under a unified experimental framework rather than to serve as a direct numerical comparison with previously published benchmark results.

8 Limitations

Although the proposed pipeline significantly reduces human involvement in the data creation and annotation process, a major limitation is its high computational cost and time requirements. The multiple stages involved in the pipeline, including data generation, data lemmatization, sense augmentation, and model fine-tuning, require substantial computational resources and extended training time. This can make the approach challenging to scale, particularly for experiments involving larger datasets and models.

Additionally, the effectiveness of the approach depends on the quality and accuracy of the automatically generated sense-augmented data. Errors or inconsistencies introduced during this stage may propagate into the fine-tuning process and affect the final translation performance.

Another limitation of work is that, since we had different data splits as compared to the previous works by Aralikatte et al. (2021) and Maheshwari et al. (2024), we are unable to do a direct comparison of our BLEU and chrF scores with theirs.

9 Conclusion and Future Work

MT models and LLMs have achieved significant progress, but, continue to struggle with capturing fine-grained semantic nuances, particularly in resource-constrained and

linguistically rich languages such as Sanskrit. Despite the use of contextual embeddings, models often fail to assign the correct sense of ambiguous words based on their context, leading to semantically inaccurate translations. In this work, we demonstrate that sense-augmented fine-tuning can effectively address this limitation by providing explicit sense-level information during training. Our experiments show that sense augmentation consistently improves translation quality by enhancing lexical disambiguation across different model architectures. While translation-oriented models such as Sarvam already perform strongly with parallel corpus fine-tuning, sense augmentation provides further improvements, and enables general-purpose LLMs such as Llama 3.1 8B to achieve performance comparable to dedicated translation models. These results highlight the potential of sense-augmented training as a complementary approach for improving semantic accuracy in MT, especially for low-resource and semantically complex languages.

As future work, we aim to explore whether our sense inventory can be made more coarse-grained to achieve a better balance between semantic specificity and effective model learning. We also plan to investigate additional information available in the MW dictionary that could contribute to improved translation and disambiguation performance, and incorporate relevant features into the sense-augmented corpus. Furthermore, we intend to evaluate whether integrating a dedicated WSD model as a separate pipeline component with existing MT models can achieve comparable performance on this dataset.

Furthermore, we plan to utilize other datasets, such as Sāmayik, within the same experimental framework to perform a comparative evaluation against the Itihāsa dataset. We also intend to evaluate models on a combined dataset comprising both Itihāsa and Sāmayik to investigate whether integrating diverse data sources improves generalization and overall performance.

10 Acknowledgment

We would like to thank Gouri Desai for her help with the data evaluation.

Appendix A Prompt Templates

1. The following prompt template was used for sense extraction with Llama:

Sense Extraction Prompt

You are a Sanskrit linguistics expert tasked with identifying the precise sense of Sanskrit words based on their English translations.

CONTEXT:
A Sanskrit text has been translated into English. Each Sanskrit word (lemma) has multiple possible meanings or senses, numbered in a list. Your task is to determine which numbered sense is used in the translation.

INSTRUCTIONS:
1. Carefully read the English translation and understand the context.
2. Examine each of the numbered senses for the lemma.
3. Consider:
- Which sense best matches how the word is used in the translation
- The cultural and philosophical context if relevant
- Whether the sense is literal or figurative in this context
- The semantic field of surrounding words in the translation
4. Determine which ONE numbered sense best matches how the word is used in the translation.
5. Provide brief reasoning for your selection.
6. If there is no correspondence between the lemma, its possible senses and the given translation, set "Selected sense" to null.

2. The following prompt template was used for instruction fine-tuning Llama and Sarvam for translation.

Translation Prompt

You are an expert Sanskrit-to-English translator. Translate the Sanskrit verse into English. Provide only the English translation.

References

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., Goldberg, Y. (2016). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*, ,

Aralikatte, R., de Lhoneux, M., Kunchukuttan, A., Søgaard, A. (2021, August). Itihasa: A large-scale corpus for Sanskrit to English translation. T. Nakazawa et al. (Eds.), *Proceedings of the 8th workshop on asian translation (wat2021)* (pp.

191–197). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.wat-1.22/

Aralikatte, R., Gantayat, N., Panwar, N., Sankaran, A., Mani, S. (2018). Sanskrit sandhi splitting using seq2 (seq) 2. *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 4909–4914).

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171–4186).

Gérard, H. (2003). Lexicon-directed segmentation and tagging of sanskrit. *Xiith world sanskrit conference, helsinki, finland, aug* (pp. 307–325).

Hellwig, O. (2017). Coarse semantic classification of rare nouns using cross-lingual data and recurrent neural networks. C. Gardent & C. Retoré (Eds.), *Proceedings of the 12th international conference on computational semantics (IWCS) — long papers.* Retrieved from https://aclanthology.org/W17-6811/

Hellwig, O., & Biagetti, E. (2025). The sanskrit sembank: O. hellwig, e. biagetti. *Language Resources and Evaluation*, 1–24,

Hellwig, O., Sellmer, S., Nehrdich, S., Widmer, P., Sennrich, R. (2026, 05). Sanskrit word sense disambiguation based on lexicographic definitions..

Itankar, P.Y., & Raza, N. (2020). Ambiguity resolution: An analytical study.

Jurafsky, D., & Martin, J.H. (2026). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition, with language models* (3rd ed.). Retrieved from https://web.stanford.edu/ jurafsky/slp3/ (Online manuscript released January 6, 2026)

Koul, N., & Manvi, S.S. (2021). A proposed model for neural machine translation of sanskrit into english. *International Journal of Information Technology*, *13*(1), 375–381,

Krishnan, S., Kulkarni, A., Huet, G. (2025). Normalized dataset for sanskrit word segmentation and morphological parsing. *Language Resources and Evaluation*, *59*(2), 1279–1330,

Maheshwari, A., Gupta, A., Krishna, A., Singh, A.K., Ramakrishnan, G., Gourishetty, A.K., Singla, J. (2024, May). Samayik: A benchmark and dataset for English-Sanskrit translation. N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 14298–14304). Torino, Italia: ELRA and ICCL. Retrieved from https://aclanthology.org/2024.lrec-main.1245/

Mikolov, T., Yih, W.-t., Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746–751).

Mukherjee, A., & Shrivastava, M. (2023, December). MEE4 and XLsim : IIIT HYD's submissions' for WMT23 metrics shared task. P. Koehn, B. Haddow, T. Kocmi, & C. Monz (Eds.), *Proceedings of the eighth conference on machine translation* (pp. 800–805). Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.wmt-1.66/

Mukherjee, A., & Shrivastava, M. (2024, November). chrF-S: Semantics is all you need. B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the ninth conference on machine translation* (pp. 470–474). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.wmt-1.33/

Nehrdich, S., Hellwig, O., Keutzer, K. (2024, November). One model is all you need: ByT5-Sanskrit, a unified model for Sanskrit NLP tasks. Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 13742–13751). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.findings-emnlp.805/

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).

Patel, D.K., & Kulkarni, A. (2024, February). Word sense alignment of Sanskrit lexica. A. Bhattacharya (Ed.), *Proceedings of the 7th international sanskrit computational linguistics symposium* (pp. 1–13). Auroville, Puducherry, India: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.iscls-1.1/

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. *Proceedings of the tenth workshop on statistical machine translation* (pp. 392–395).

Pradeep, A., Krishnan, S., Mamidi, R. (2026, March). Towards building a computational sense inventory from the monier-Williams dictionary using clustering techniques. P. Satuluri & P. Goyal (Eds.), *Proceedings of the 8th international Sanskrit computational linguistics symposium* (pp. 32–46). IIT Roorkee, Roorkee, India: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2026.iscls-1.3/

Pradeep, A., & Mamidi, R. (2025). Sandarśana: A survey on sanskrit computational linguistics and digital infrastructure for sanskrit. *ACM Computing Surveys*, *57*(10), , https://doi.org/10.1145/3729530 Retrieved from https://doi.org/10.1145/3729530

Pradeep, A., Mamidi, R., Satuluri, P. (2024, February). Context and WSD: Analysing Google Translate's Sanskrit to English output of bhagavadgītā verses for word meaning. A. Bhattacharya (Ed.), *Proceedings of the 7th international sanskrit computational linguistics symposium* (pp. 14–26). Auroville, Puducherry, India: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.iscls-1.2/

Punia, R., Sharma, A., Pruthi, S., Jain, M. (2020, December). Improving neural machine translation for Sanskrit-English. P. Bhattacharyya, D.M. Sharma, & R. Sangal (Eds.), *Proceedings of the 17th international conference on natural language processing (icon)* (pp. 234–238). Indian Institute of Technology Patna, Patna, India: NLP Association of India (NLPAI). Retrieved from https://aclanthology.org/2020.icon-main.30/

Rei, R., Stewart, C., Farinha, A.C., Lavie, A. (2020, November). COMET: A neural framework for MT evaluation. B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (emnlp)* (pp. 2685–2702). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.emnlp-main.213/

Riemer, N. (2010). *Introducing semantics*. Cambridge university press.

Sitender, & Bawa, S. (2021). A sanskrit-to-english machine translation using hybridization of direct and rule-based approach. *Neural Computing and Applications*, *33*(7), 2819–2838,

Venkatesh, T.S.P., Shetye, K.M., Arjunasamy, V., Sahu, A.K., Krishnan, S., Krishnamurthy, P. (2026, March). Santham: A curated Sanskrit–Tamil dataset with anvaya and segmentation for building and evaluating machine translation. P. Satuluri & P. Goyal (Eds.), *Proceedings of the 8th international Sanskrit computational linguistics symposium* (pp. 65–80). IIT Roorkee, Roorkee, India: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2026.iscls-1.5/