MaitH 1.0: A Parallel Corpus and Baseline for Low-Resource Maithili-Hindi Translation

Kamanksha Prasad Dubey[1], **Chandresh Kumar Maurya**[1], **Kumar Padmanabh**[2]
[1]Indian Institute of Technology Indore, Simrol, Indore, Madhya Pradesh, India 453552
[2]EBTIC Khalifa University, Abu Dhabi, UAE
{phd2201101001, chandresh}@iiti.ac.in, kumar.padmanabh@ku.ac.ae

Abstract

Maithili is one of the 22 official languages recognized in the Indian Constitution. The literature of Maithili is rich; however, due to current socio-political changes, the language is on the verge of extinction. Therefore, it is crucial to develop a corpus for low-resource Indic languages like Maithili to ensure that the dream of "No Language Left Behind" (NLLB) is realized. With this in mind, we contribute a corpus (1,05,600 sentences) containing both manually curated and synthetically generated. Additionally, we propose a strong baseline on the Maithili-Hindi pair using multilingual pretrained models such as IndicTrans2, mBART50, mT5, and NLLB-200 distilled. We evaluate the translation systems using standard performance metrics, including BLEU, CHRF2, TER, COMET, METEOR, and BERTScore. Comparative experiments conducted against the existing NLLB dataset (550,300 sentence pairs) demonstrate that our proposed dataset consistently yields superior translation quality. Finally, these results demonstrate that, even with a smaller corpus size, high-quality, task-specific data significantly enhance translation accuracy for low-resource Indian languages, such as Maithili. To support reproducibility and further research in Maithili to Hindi machine translation, we publicly release our Maithili-Hindi parallel dataset. The dataset and code are publicly available at https://github.com/kamanksha-prog/MaitH_1.0/

Keywords: Low-Resource Language, Maithili, Hindi, Neural Machine Translation

1. Introduction

Machine translation (MT) has witnessed significant advancements over the past decade, driven largely by the availability of extensive parallel corpora and sophisticated models. However, these advancements are predominantly focused on high-resource languages, leaving many low-resource languages with limited or no effective translation systems. Maithili, a language spoken by over 22M people (according to Wiki[1]) primarily in the eastern regions of India and the southern plains of Nepal, is one such low-resource language. Despite its rich linguistic heritage and substantial speaker base, Maithili remains underrepresented in the realm of natural language processing (NLP), particularly in machine translation.

The development of effective translation systems for low-resource languages like Maithili is crucial for several reasons. First, it helps in preserving linguistic diversity by enabling communication between speakers of different languages. Second, it provides access to information and services for speakers of these languages, contributing to social and economic inclusion. Finally, it adds to the global corpus of linguistic data, which is essential for studying and understanding human languages.

Several studies have focused on building translation systems for Indian languages, particularly those with limited resources. INDICNLP Project (Kunchukuttan, 2020) is a notable initiative to develop NLP resources and tools for Indian languages. It includes datasets, word embeddings, and other linguistic resources for multiple Indian languages, including low-resource languages. Researchers have created bilingual and multilingual corpora for Indian languages, which serve as essential resources for training translation models. For instance, (Kunchukuttan et al., 2018) developed the IIT Bombay Hindi-English corpus, a significant resource for Hindi-English translation tasks. There have been attempts to develop corpora and build translation systems for specific regional languages in India, such as (Post et al., 2013; Revanuru et al., 2017; Laskar et al., 2019, 2020; Pathak et al., 2019; Pathak and Pakray, 2019; Choudhary et al., 2018; Singh et al., 2018). However, similar efforts for Maithili remain sparse.

The No Language Left Behind (NLLB) (Costa-Jussà et al., 2022; Tiedemann, 2012) dataset is a part of Meta AI's NLLB initiative, which aims to improve machine translation for low-resource languages. The dataset includes parallel text for 200+ languages, including Maithili-Hindi. It is constructed from multiple sources, such as web-crawled data and publicly available datasets. There are 5,50,300 Maithili-Hindi parallel sentences available in OPUS[2] (Tiedemann, 2012; Fan et al., 2021; Schwenk et al., 2019). Furthermore, while the NLLB dataset provides a large number of Maithili-Hindi parallel sentences, its quality is poorer due to

[1]https://en.wikipedia.org/wiki/Maithili_language

[2]http://opus.nlpl.eu

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8567–8576
11–16 May 2026.

automatically generated translations and misalignments. In contrast, our dataset, although smaller (1,05,600 sentences), includes 5,600 manually verified sentences, and the rest are synthetically generated. Further comparison of our data with NLLB is provided in the experiment section.

Our contributions in the paper are as follows:

- We contribute a Maithili to Hindi parallel corpus comprising 105,600 sentences, which includes 5,600 manually verified sentences, and the remaining 100,000 sentences are synthetically generated.
- We fine-tune the SOTA MT models to present a strong baseline of Maithili to Hindi translation task and show the superior quality of our data compared to the NLLB dataset.

2. Corpus Creation Methodology

We construct our corpora by using web scraping and optical character recognition (OCR) techniques. Data is sourced from various online repositories and printed materials, with different domains as detailed in Table 1. Web scraping is done on four websites: khattarkaka[3], videhamaithili[4], pranaw jha's blogs[5], and maithilijindabaad[6]. Additionally, OCR is used for 141 books selected from the Maithili books collection[7]. This section outlines the steps involved in creating the Maithili monolingual corpus and the Maithili to Hindi parallel dataset, as illustrated in Figure 1. Furthermore, we discuss the creation of manual and synthetic parallel datasets in detail, along with their quality checks.

2.1. Book Digitization for Corpus Development

The Maithili data is collected from PDF files of various genres of books, such as stories, conversations, and articles, for our Maithili to Hindi MT task. We use Python libraries to extract the text and then process it. Specifically, we extract Maithili text from a PDF using Tesseract OCR (pytesseract)[8] in Python; this process involves converting PDF pages to images and then applying OCR to extract text from those images. To our understanding, Tesseract does not have a dedicated Maithili language model. However, Maithili uses the Devanagari script, which is supported by Tesseract's Hindi (hin) language data. This allows Tesseract to recognize Maithili text using the Hindi model. There may be a recognition mistake by OCR, which we did not handle in the present study.

S.N	Sources	Domain
1	khattarkaka	story, novel, satire
2	videhamaithili	literature, culture, history, society
3	pranawjha.blogs	articles, story
4	maithilijindabaad	literature, philosophy, culture, heritage, news
5	maithili-books	literature, story, history, culture.

Table 1: List of resources used to extract the monolingual Maithili corpus.

2.2. Automated Web Scraping

We collect the Maithili data through web scraping using Selenium (Gojare et al., 2015) and BeautifulSoup[9] in Python. Because several webpages load content dynamically, we first use Selenium to render the complete page and obtain the full HTML source. We then parse the rendered content with BeautifulSoup and extract textual material primarily from heading and paragraph elements. During extraction, we exclude non-content sections such as navigation menus, advertisements, and hyperlinks. When container elements (e.g.,) contain meaningful textual content, we retain their text. After extraction, we remove all HTML tags and clean the text by eliminating encoding artifacts, invisible Unicode characters (e.g., U+200B), and unnecessary whitespace, while preserving valid Devanagari script content.

2.3. Data Cleaning

Once we extract the text, regex scripts are employed for text processing to remove English text and format the Maithili content. This involves removing unnecessary characters or symbols, normalizing the text (redundant punctuation marks, non-printable characters, and extra spaces), and segmenting the text into sentences. The purpose of writing regex code is to clean the data as much as possible and make it structured. The cleaned Maithili data (1,00,000 sentences) is then stored in a text file format.

[3] https://khattarkaka.com
[4] https://videhamaithili.wordpress.com/
[5] http://pranawjha.blogspot.com
[6] https://maithilijindabaad.com
[7] https://archive.org/details/432-MAITHILI-BOOKS
[8] https://github.com/madmaze/pytesseract
[9] https://pypi.org/project/beautifulsoup4

Dataset	Language	Sentences	Tokens	Mean	Median	S.D	TTR	Replaced by <unk>
MaitH 1.0 (Our Dataset)								
Train	Maithili	84,480	2,192,627	16.25	13	17.63	0.0559	0.127%
	Hindi	84,480	2,130,752	19.19	15	17.66	0.0576	0.0023%
Dev	Maithili	10,560	269,184	15.79	13	15.52	0.4558	0.0858%
	Hindi	10,560	262,342	19.15	15	17.52	0.4562	0.00191%
Test	Maithili	10,560	253,779	15.90	11	12.58	0.4835	0.0686%
	Hindi	10,560	236,349	18.60	14	13.75	0.5190	0.000846%
NLLB Maithili–Hindi Dataset								
Train	Maithili	440,240	5,328,862	6.53	5	4.34	0.0230	0.02%
	Hindi	440,240	3,745,318	5.78	5	3.23	0.0327	0.409%
Dev	Maithili	55,030	682,981	6.66	6	4.38	0.1796	0.0171%
	Hindi	55,030	470,957	5.77	5	3.12	0.2604	0.43%
Test	Maithili	55,030	687,058	6.69	6	4.44	0.1785	0.0192%
	Hindi	55,030	473,530	5.79	5	3.12	0.2590	0.42%

Table 2: Comparative statistics of the MaitH 1.0 and existing NLLB Maithili–Hindi parallel datasets. Mean, Median and S.D (Standard Deviation) refer to the number of tokens in sentences, TTR = Type–Token Ratio, Replaced by <unk> = Percentage of tokens replaced by unknown word.

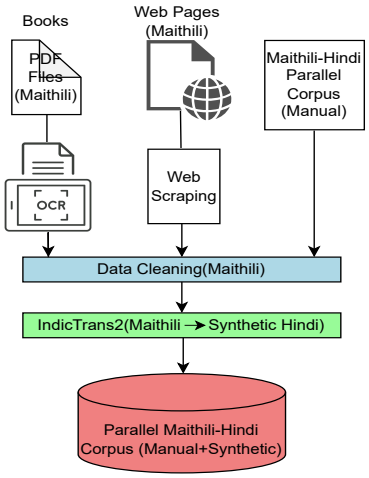

Figure 1: Detailed workflow for creating the Maithili monolingual corpus and Maithili-Hindi parallel dataset.

2.4. Manual and Synthetic data Generation

For manual translation, we gathered 5,600 Maithili texts from khattarkaka, pranawjha.blogs, and reviewed them thoroughly. Two linguistic experts, proficient in both Maithili and Hindi languages, translated the 5,600 Maithili text sentence by sentence, ensuring accuracy and coherence.

For pseudo data generation, we collected the Maithili corpus from videhamaithili, maithilijindabad, and maithili-books, then applied data augmentation (Sennrich et al., 2016) techniques to address the limited parallel corpora for Maithili-Hindi. IndicTrans2 (Gala et al., 2023) is used to generate synthetic parallel data by translating 1,00,000 Maithili monolingual corpus into Hindi. These synthetic sentences are paired with their original Maithili counterparts to create additional parallel sentence pairs. The overall 1,05,600 sentences increase the training data size, exposing the NMT models to more diverse sentence structures and vocabulary.

2.5. Maithili-Hindi parallel dataset Statistics

Table 2 presents the dataset statistics, including the number of sentences, tokens, type-token ratio (TTR), percentage of tokens replaced by <unk>, average sentence lengths (in tokens), median, and S.D. (standard deviation) for the Maithili and Hindi datasets across the train, development, and test splits. The descriptive details of the existing NLLB corpus of the Maithili to Hindi parallel sentence pair are shown in the part of the same Table 2.

2.5.1. Comparative Analysis with Existing Dataset

Compared to the NLLB Maithili–Hindi dataset, MaitH 1.0 demonstrates several distinct advantages that make it more suitable for robust machine translation research, as summarized in Table 2. First, the mean sentence length in MaitH 1.0 is approximately 16–19 tokens, which is nearly three times longer than that of NLLB, around 6–7 tokens, providing richer syntactic and semantic context for model learning. Second, the higher type-token ratio (TTR) values in MaitH 1.0 (0.056-0.51) indicate greater lexical diversity, which helps translation models capture morphological and vocabulary variations more effectively. In contrast, NLLB exhibits much lower TTR values, i.e., 0.023 to 0.26, suggesting a simpler and more repetitive vocabulary. Third, MaitH 1.0 shows minimal token replacement by unknown word (<unk>), i.e., below 0.13%, reflecting comprehensive vocabulary coverage and

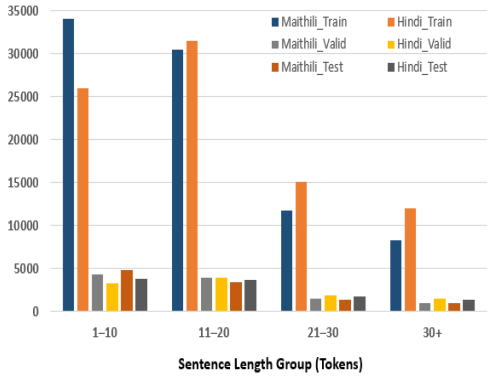

Figure 2: Grouped sentence length distribution for MaitH 1.0 dataset.

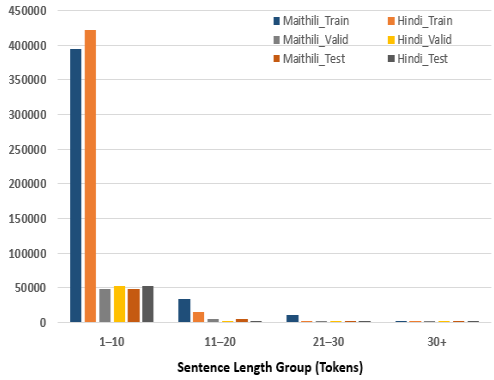

Figure 3: Grouped sentence length distribution for NLLB dataset.

cleaner tokenization, while the NLLB dataset has comparatively higher ‹unk› rates than our dataset.

Furthermore, the higher standard deviation values in Maith 1.0 (12.58-17.66) quantify the wider range of sentence length, which includes both short and long sentences, and introduces structural diversity, which is essential for developing generalizable models. In contrast, NLLB has a smaller S.D (3.12-4.44), showing sentences are shorter and uniform in length.

In general, these characteristics make MaitH 1.0 a richer, more challenging, and linguistically diverse data set that better represents the use of Maithili-Hindi in the real world compared to the existing NLLB corpora.

2.5.2. Sentence Length Distribution Analysis

To assess the structural characteristics of the parallel corpora, we performed a detailed sentence length distribution analysis for both the MaitH 1.0 and NLLB datasets. Sentences were grouped into four token length ranges: 1-10, 11-20, 21-30, and 30+.

As shown in Figure 2 and Figure 3, the MaitH 1.0 dataset exhibits a balanced distribution across these groups, with a substantial proportion of medium and long sentences (11–30 and 30+ tokens) for both Hindi and Maithili splits. This indicates that the corpus predominantly consists of medium-length sentences. In contrast, the NLLB dataset shows a heavy skew toward shorter sentences (1–10 tokens), suggesting that its examples are more fragmented or simple in structure. The grouped sentence length distribution highlights that MaitH 1.0 provides greater lexical and syntactic diversity compared to NLLB, thereby offering richer contextual information for model learning.

2.6. Automatic Quality Check

To assess the quality and alignment of our parallel datasets, we employ two SOTA multilingual sentence embedding models: Language-agnostic BERT Sentence Embedding (LaBSE) (Feng et al., 2022) and LASER (Artetxe and Schwenk, 2019). These models project sentences from different languages into a shared semantic space, enabling direct comparison through cosine similarity. Table 3 provides a comparative analysis, showcasing the average similarity scores and variance for each dataset using both LaBSE and LASER.

From the results in Table 3, the manually curated parallel corpus exhibits the highest average similarity and the lowest standard deviation, indicating consistently strong alignment and minimal noise. In contrast, pseudo-parallel and NLLB data show comparatively lower average similarity and higher variability, reflecting weaker alignment and greater heterogeneity. We note that pseudo-parallel data has not been validated manually, possibly due to its large size and the availability of an expert in the Maitahli language. However, we show, in the experiment, that the baseline models achieve the highest performance on the *combined data* compared to manually created data alone, indicating the value of pseudo-parallel data. These findings underscore the importance of high-quality human-aligned data for building robust multilingual models and also provide an objective basis for selecting or filtering parallel corpora for low-resource machine translation tasks.

3. Experiments

This section presents data preprocessing, baseline models, evaluation metrics, results, and discussions. For the experiment, all the datasets are divided in the ratio of 80/10/10 for train/valid/test unless otherwise mentioned.

3.1. Data Preprocessing

The raw data often contains inconsistencies in text formatting, including varying Unicode encodings

Dataset	Sentences	Median (sent. sim.)	S.D (sent. sim.)	LaBSE	LASER2
Manually Created	5,600	**0.7129**	**0.1660**	**0.6925**	**0.7265**
Pseudo-Parallel	1,00,000	0.6952	0.1927	0.6678	0.4815
Combined (Manually + Pseudo)	1,05,600	0.6963	0.1915	0.6691	0.5026
NLLB	5,50,300	0.6958	0.2086	0.6659	0.3779

Table 3: Comparison of sentence embedding similarity across datasets. "sent. sim." denotes cosine similarity between Maithili–Hindi sentence pairs. LaBSE and LASER2 represent multilingual embedding models used to compute similarity.

and the use of non-standard characters. We standardize the text by converting all characters to their normalized forms using Unicode normalization and applying the standard IndicNLP normalization (Kunchukuttan, 2020) to the corpus. The pretrained SentencePiece Model (SPM) (Gala et al., 2023) is used for subword tokenization (Kudo and Richardson, 2018). SentencePiece is an unsupervised subword tokenizer that efficiently handles the morphological richness of Maithili and Hindi. The final dictionaries for Maithili and Hindi comprised 1,22,706 and 1,22,672 unique subword units, respectively.

3.2. Baseline Models

We finetune four pre-trained multilingual models as baselines on MaitH 1.0 and NLLB training dataset: IndicTrans2 (Gala et al., 2023), mT5 (Xue et al., 2021), mBART50 (Liu et al., 2020), and NLLB-200 distilled model[10]. Each model is trained using task-specific hyperparameter configurations tailored for low-resource neural machine translation. In this section, we provide the details of the hyperparameters used for finetuning the models to ensure clarity and reproducibility.

We finetune the IndicTrans2 custom 12-layer transformer with 512 embedding dimensions using Adam optimizer (Kingma and Ba, 2014) ($\beta(0.9, 0.98)$), a 3×10^{-5} learning rate, 0.1 label smoothing, and an inverse square root scheduler with 2000 warmup updates. Training runs for 35 epochs on an NVIDIA RTX A5000 24GB VRAM GPU with a 0.2 dropout rate, gradient clipping (norm 1.0), mixed precision (fp16), and a maximum batch size of 2,048 tokens.

mT5 (Xue et al., 2021) model comprises 12 encoder and 12 decoder layers, with 12 attention heads and an embedding dimension of 768. The feed-forward network (FFN) dimension is set to 2048, using a GeGLU activation function, a linear learning rate scheduler, a learning rate of 5×10^{-5}, with a dropout rate of 0.1. The model trains for 7 epochs with a batch size of 4 on an NVIDIA RTX A4500 20GB VRAM GPU.

[10] https://huggingface.co/facebook/nllb-200-distilled-600M

In our experiment, we finetune the pretrained *mBART50* model (Liu et al., 2020) with 12 encoder and decoder layers, each comprising 16 attention heads and an embedding dimension of 1024. The feed-forward network (FFN) dimensions are set to 4096. A dropout of 0.1, a learning rate of 5×10^{-5}, and the activation function uses ReLU. The training runs for 7 epochs with a batch size of 6, conducted on the same machine as mT5.

The NLLB-200 distilled model is finetune using the Hugging Face transformers library on a Maithili–Hindi parallel corpus. It is initialized from a publicly available distilled NLLB-200 checkpoint provided by Meta AI. The model follows the M2M100ForConditionalGeneration architecture with 12 encoder and 12 decoder layers, 16 attention heads, a hidden size of 1024, and a feed-forward dimension of 4096. All input and output sequences are truncated and padded to 128 tokens. The model is trained on an NVIDIA RTX A5000 GPU for 5 epochs with a learning rate of 2×10^{-5}, batch size of 8, weight decay of 0.01, and the learning rate scheduler is set to linear. Mixed-precision (FP16) training is enabled, and the best checkpoint is selected based on the lowest evaluation loss.

3.3. Evaluation Metrics

We report well-known evaluation metrics such as BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), character-level precision-recall F-score (ChrF2) (Popović, 2015), Crosslingual Optimized Metric for Evaluation of Translation (COMET) (Rei et al., 2020), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2019), and Translation EDIT Rate (TER) (Snover et al., 2006) metrics. The higher is the better for the first five metrics, whereas a lower value for TER is preferred. In addition, we also report the human evaluation metrics, i.e., adequacy and fluency scores, by two linguistic experts of Maithili to Hindi translation. Adequacy measures the meaning of the source sentence that is preserved in the target translation, while fluency measures the grammatical correctness and naturalness of the translated text in the target language. The adequacy and fluency of translations are evaluated on a 4-point scale ranging from 1 to 4. The details of human evaluation are discussed in the section 3.4.3.

Model	BLEU	chrF2	TER	COMET	METEOR	BERTScore
IndicTrans2	4.01	21.54	0.97	0.4365	0.2036	0.8835
mT5	26.56	54.62	0.58	0.6903	0.5000	0.9354
mBART50	23.82	52.90	0.61	0.6740	0.4850	0.9306
NLLB-200	**28.57**	**57.15**	**0.55**	**0.7292**	**0.5277**	**0.9387**

Table 4: Results of the models trained and tested on the manually created data

Model	Training Data	BLEU	chrF2	TER	COMET	METEOR	BERTScore
IndicTrans2	Our	9.60	32.91	0.86	0.5248	0.3206	0.8951
	NLLB	2.12	16.41	1.40	0.4291	0.1353	0.8631
mT5	Our	15.42	47.38	0.71	0.5814	0.4614	0.9142
	NLLB	10.44	34.61	1.01	0.5679	0.2646	0.8835
mBART50	Our	25.94	52.85	0.63	0.6711	0.4865	0.9223
	NLLB	8.12	28.95	1.26	0.5181	0.1895	0.8732
NLLB-200	Our	**37.97**	**59.90**	**0.55**	**0.7356**	**0.5644**	**0.9382**
	NLLB	13.86	35.16	1.27	0.5747	0.2489	0.8802

Table 5: Each model is trained separately on our (MaitH 1.0) dataset and the NLLB dataset, and evaluated on the MaitH 1.0 test set.

We use sacrebleu[11] library to compute BLEU and chrF scores, pyter[12] library to compute TER score, and Unbabel[13] library to compute COMET, bert_score[14] library to compute BERTScore, meteor_score[15] library is used to compute METEOR, which is a standard for evaluating machine translation outputs.

3.4. Results and Discussions

This section presents results from baseline models for the Maithili→ Hindi translation direction. We first discuss the result on *only* manually curated data and then on the combined data, while also comparing the results obtained on the NLLB data. Experimental results are presented in tables 4 and 5.

From the table 4, we can observe that NLLB-200 outperforms the other three models on all metrics, likely due to it being pretrained on a massive amount of *parallel data* and a large number of languages, helping cross-lingual transferability. Comparing metrics in Tables 5, we can see that models trained on the MaitH 1.0 consistently outperform models trained on the NLLB training data. It shows the superior quality of the MaitH 1.0 dataset. Comparing the results in Tables 4 and 5, we can see that pseudo-parallel data further improves the performance metrics, thus supporting the value addition by pseudo-parallel data. Moreover, we observe that it is not true for mBART50 and mT5, likely because these two models do not include Maithili in their supported languages, whereas IndicTrans2 and NLLB have explicit coverage for Maithili, enabling them to handle the translation task better.

3.4.1. Zero-shot Evaluation of Pretrained Models

To better understand the contribution of fine-tuning on our data, we also evaluate the pretrained models in a zero-shot setting, on the MaitH 1.0 test (10,560) dataset to establish baseline performance without any task-specific training. As shown in Table 6, all models perform poorly without fine-tuning (refer to results in table 5 and 4), indicating limited cross-lingual transfer for this low-resource pair.

3.4.2. Model Output on Test Samples

To analyze the performance of our fine-tuned models, we present sample translations from our Maithili to Hindi test dataset. Figure 6 below showcases translations generated by IndicTrans2, mBART50, mT5, and NLLB-200, alongside the original Maithili sentence and the reference Hindi translation. The comparison highlights the differences in translation quality across models.

3.4.3. Human Evaluation

The human study focused on the adequacy and fluency scores. We conducted a human evaluation on a randomly selected set of 200 Maithili to Hindi system translations. The evaluation involves rating each target translation against a reference translation using a scale of 1-4, with 4, 3, 2, and

[11] https://pypi.org/project/sacrebleu
[12] https://pypi.org/project/pyter3
[13] https://github.com/Unbabel/COMET
[14] https://github.com/Tiiiger/bert_score
[15] https://www.nltk.org/api/nltk.translate.meteor_score.html

Model	BLEU	ChrF2	TER	COMET	METEOR	BERTScore
IndicTrans2	–	–	–	–	–	–
mT5	4.68	29.87	0.93	0.4510	0.1759	0.8651
mBART50	3.98	28.83	0.93	0.4337	0.1693	0.8618
NLLB-200	16.34	40.57	0.90	0.6092	0.3043	0.8959

Table 6: Zero-shot Maithili to Hindi translation results of pretrained multilingual models on the MaitH 1.0 (test) dataset.

Model	Adequacy		Fluency	
	Mean	SD	Mean	SD
IndicTrans2	1.92	1.01	1.80	0.97
mT5	2.64	1.24	2.32	1.20
mBART50	3.05	1.09	2.90	1.12
NLLB-200	3.18	1.08	3.00	1.10

Table 7: Mean and standard deviation (SD) of adequacy and fluency scores across models.

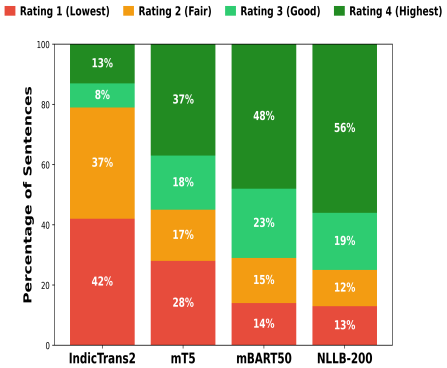

Figure 4: Percentage distribution of Adequacy ratings (1–4 scale).

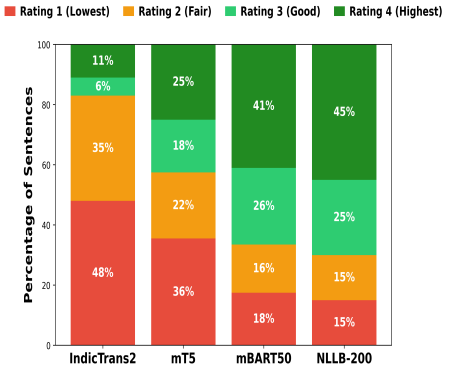

Figure 5: Percentage distribution of Fluency ratings (1–4 scale).

1 signifying the highest quality, good, fair, and the lowest. Figure 4 and 5 show the percentage distribution of adequacy and fluency ratings out of randomly chosen 200 Maithili to Hindi translation sentences of each model. Then we calculate the average adequacy and fluency score after manually rating them and show it in Table 7. Moreover, we have also evaluated the standard deviation score of adequacy and fluency for each model. The NLLB-200 model attains the highest adequacy (3.18) and fluency (3.00) scores, followed by mBART50, whereas mT5 obtained moderate scores. The IndicTrans2 records the lowest performance compared to all other models.

4. Conclusion and Future works

Our work contributes a manually curated and synthetically generated parallel corpus for the Maithili-Hindi language pair. We also develop a strong baseline for Maithili-Hindi translation using our dataset. The study reveals the value of manually created and validated data (compared against NLLB, which is noisy). Future work will focus on improving synthetic data quality and incorporating domain-specific data. Additionally, large language models or a new architecture with Maithili and Hindi-specific linguistic features may enhance their capabilities.

5. Limitations

This study, despite its valuable insights, faces limitations due to limited manually curated data and reliance on synthetic data. The small dataset size and potential noise in synthetic data hinder model performance. More robust validation is needed for synthetic data to improve its quality. Improved training procedure in a low-data regime may help address these limitations and improve translation accuracy and fluency. Further, we agree that using OCR may cause errors during data collection, but we have not handled the OCR error explicitly. Data quality can be further improved by manual inspection in a post-editing process, which we have left due to cost. The LaBSE/LASER similarity scores are not used to filter out noisy or low-quality sentence triplets. Thus, data quality may be further improved by filtering out low-score sentences.

6. Ethical considerations

This research on Maithili to Hindi machine translation adhered to ethical principles, including data

Source 1	हमरा मुँह सँ किछु उत्तर नहि बहराएल ।
Reference translation	मेरे मुँह से कुछ उत्तर नहीं निकला।
Gloss:	No answer came out of my mouth.
IndicTrans2	मुझे मुँह से कुछ जवाब नहीं मिला।
Gloss:	I did not get any reply from my mouth.
mT5	मेरे मुँह से कुछ उत्तर नहीं बहराया।
Gloss:	No reply came out of my mouth.
mBART50	मेरे मुँह से कुछ भी उत्तर नहीं निकला।
Gloss:	No reply came out of my mouth.
NLLB-200	मेरे मुँह से कोई जवाब नहीं निकला।
Gloss:	No answer came out of my mouth.
Source 2	आब जा कऽ बड़की बाबी ओ सहजोपीसी कैं वस्तुस्थितिक बोध भेलैन्ह ।
Reference translation	अब जाकर बड़ी दादी और सहजो बुआ को वस्तुस्थिति का बोध हुआ।
Gloss:	Now elder grandmother and Sahajo Bua realized the reality.
IndicTrans2	अब जब बड़ी बड़ी दादी और सहार को वस्तु की वस्तुएँ हुई थीं।
Gloss:	Now when the elder grandmother and Sahar had become objects of the matter.
mT5	अब जाकर बड़ी दादी और सहजोपीसी को वस्तुस्थिति का बोध हुआ।
Gloss:	Now the elder grandmother and sister-in-law realized the true situation.
mBART50	अब जाकर बड़ी दादी और सहजो बुआ को वस्तुस्थिति का बोध हुआ।
Gloss:	Now elder grandmother and Sahajo Bua realized the reality.
NLLB-200	अब जाकर बड़ी दादी और सहजोपीसी को वस्तुस्थिति का एहसास हुआ।
Gloss:	It was only now that the elder grandmother and Sahajopisi realized the factual situation.
Source 3	हम चुपचाप अपन सूटकेस ओ विस्तर उठाओल और रिक्सापर चढि पराजित सैनिक जकाँ स्टेशन विदा भेलहुँ।
Reference translation	मैं चुपचाप अपना सूटकेस और बिस्तर उठाया और रिक्शा पर चढ़ कर पराजित सैनिक जैसे स्टेशन विदा हुआ ।
Gloss:	I silently picked up my suitcase and bedding, boarded a rickshaw and left the station like a defeated soldier.
IndicTrans2	मैं चुपचाप चुपचाप अपने तीरों को उठा और दरवाजे पर चढ़ाई की तरह चढ़ गया।
Gloss:	I quietly and silently picked up my arrows and climbed like a ladder to the door.
mT5	मैं चुपचाप अपना सूटकेस और विस्तर उठाया और रिक्शा पर चढ़कर पराजित सैनिक की तरह स्टेशन चला गया।
Gloss:	I silently picked up my suitcase and bedding, boarded a rickshaw and went to the station like a defeated soldier.
mBART50	मैंने चुपचाप अपना सूटकेस ओ विस्तर उठाया और रिक्सा पर चढ़कर पराजित सैनिकों के जैसे स्टेशन से विदा हुआ।
Gloss:	I silently picked up my suitcase and bedding, boarded a rickshaw and left the station like a defeated soldier.
NLLB-200	मैंने चुपचाप अपना सूटकेस और चौड़ा उठाया और रिक्शे पर चढ़कर पराजित सिपाही की तरह स्टेशन से रवाना हो गया।
Gloss:	I quietly lifted my suitcase wider and climbed into the rickshaw and started from the station like a defeated soldier.
Source 4	नदी त अछिए नै! गामक बलान नदी सुखि क पीच रोड बनल अछि आ साइकिल, मोटरसाईकिल सभ ओई बाटे सरसरायल ऐ पार से ओइ पार भ रहल अछि।
Reference translation	नदी तो है ही नहीं! गांव का बलान नदी सूखकर पिच रोड बना हुआ है, और साइकिल, मोटरसाइकिल, उस रास्ते से सरसराते हुए इस पार से उसे पार हो रहे हैं।
Gloss:	There is no river! The Balan river of the village has dried up and has become a paved road, and bicycles and motorcycles are swooshing from one side to the other.
IndicTrans2	नदी तो नहीं है! गाँव के बल नदी के किनारे बने हुए हैं, साइकिल और साइकिल के पार के पार से सभी पार हो रहे हैं।
Gloss:	There is no river! The village's boats are built on the river bank, everyone is crossing it on bicycles and bikes.
mT5	नदी तो है ही नहीं! गाँव की बलान नदी सूखी पीच रोड बना हुआ है और साइकिल, मोटरसाइकिल सब उस रास्ते से सरसराए इस पार से ओइ पार हो रही है।
Gloss:	There is no river! The Balan river in the village is a dry tar road and bicycles and motorcycles are crossing it from one side to the other.
mBART50	नदी तो है ही नहीं! गाँव का बल नदी सुखकर पीच रोड बना हुआ है और साइकिल, मोटरसाइकिल सब उसी रास्ते से सरसराए हुए इस तरह से इस तरह से सरसराए जा रहे हैं।
Gloss:	There is no river! The riverbed of the village has dried up and turned into a tar road and bicycles, motorcycles, everyone is moving along that path in this way and that way.
NLLB-200	नदी तो है ही नहीं! गाँव की बलान नदी सुखी का पीच रोड बना हुआ है और साइकिल, मोटरसाइकिल सभी उसी रास्ते सरसरायल से उसी रास्ते से उसी रास्ते से गुजर रही है।
Gloss:	There is no river! The village's Balan river is a pitch road and bicycles, motorcycles, all pass through the same route from Sarasrail.

Figure 6: Sample output of the finetuned model, Maithili (source) text to Hindi (Target) translations from our test dataset. The green and red highlight the correct and incorrect or erroneous parts of translation respectively.

privacy and consent, bias and fairness, impact on low-resource languages, transparency and reproducibility, and avoidance of harm. Consent was obtained for manually curated data, and efforts were made to minimize bias in the models. The study aims to support the Maithili language community and was conducted transparently to ensure reproducibility. The research was designed to avoid any harm to individuals or communities. The code and dataset will be released under the CC BY-NC 4.0 license.

7. Bibliographical References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Himanshu Choudhary, Aditya Kumar Pathak, Rajiv Ratan Saha, and Ponnurangam Kumaraguru. 2018. Neural machine translation for English-Tamil. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 770–775, Belgium, Brussels. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Satish Gojare, Rahul Joshi, and Dhanashree Gaigaware. 2015. Analysis and design of selenium webdriver automation testing framework. *Procedia Computer Science*, 50:341–346.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Anoop Kunchukuttan. 2020. The IndicNLP Library. https://github.com/anoopkunchukuttan/indic_nlp_library/blob/master/docs/indicnlp.pdf.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The iit bombay english-hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Sahinur Rahman Laskar, Abinash Dutta, Partha Pakray, and Sivaji Bandyopadhyay. 2019. Neural machine translation: English to hindi. In *2019 IEEE Conference on Information and Communication Technology*, pages 1–6.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. 2020. Hindi-Marathi cross lingual model. In *Proceedings of the Fifth Conference on Machine Translation*, pages 396–401, 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Amarnath Pathak and Partha Pakray. 2019. Neural machine translation for indian languages. *Journal of Intelligent Systems*, 28(3):465–477.

Amarnath Pathak, Partha Pakray, and Jereemi Bentham. 2019. English–mizo machine translation using neural and statistical approaches. *Neural Comput. Appl.*, 31(11):7615–7631.

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

Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the fisher and callhome spanish–english speech translation corpus. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 138–148.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Karthik Revanuru, Kaushik Turlapaty, and Shrisha Rao. 2017. Neural machine translation of indian languages. In *Proceedings of the 10th Annual ACM India Compute Conference*, Compute '17, page 11–20, New York, NY, USA. Association for Computing Machinery.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Shivkaran Singh, M. Anand Kumar, K.P. Soman, Sabu M. Thampi, El-Sayed M. El-Alfy, Sushmita Mitra, and Ljiljana Trajkovic. 2018. Attention based english to punjabi neural machine translation. *J. Intell. Fuzzy Syst.*, 34(3):1551–1559.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.