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

We thank the reviewers for their thoughtful and constructive comments. We have carefully considered each comment and have incorporated the necessary changes to improve the quality of our manuscript. We have re-written the manuscript completely and added the new results. Below, we address each comment individually and explain the revisions made.

Reviewer [bz9c]

1. Lack of novelty, mainly focusing on corpus development.

A: We understand the reviewer's concern regarding the novelty of our work, which primarily centers on the creation of the Maithili-Hindi parallel dataset. While corpus creation itself is a significant contribution to the field of low-resource languages, we believe it serves as a critical foundation for further NLP research and applications in machine translation. We emphasize this contribution in the introduction (Section 1), detailing the challenges of working with low-resource languages and the potential impact of this dataset on future developments. **Another important point we want to make is the 'data quality'. Existing NLLB data has Maithali-Hindi pairs but is noisy and, as we showed in experiment, models achieve lower performance on it than our dataset.**

2. The monolingual data quality has not been assessed.

A: Thank you for your suggestions. We have included automatic evaluation of data quality metrics—LaBSE and LASER2 in Table 2. Due to large size and lack of availability of the expert of Maithali, we have left the manual validation of the synthetically generated dataset for future. However, we show that combining the synthetic data with manually curated dataset boots the model performance.

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

Table 2: Analysis of Avg. LaBSE, LASER, median similarity, and standard deviation across the Maithili-Hindi dataset

3. Baseline experiments present metrics without a detailed analysis.

A: Thank you for your suggestions, we have discussed our results in section 3.4, after detailed analysis we have also evaluated other semantics based metrics like COMET, METEOR, and BERTScore on our dataset and existing NLLB dataset respectively as shown in Table 3 and 4 in the paper.

Reviewer [ZB6s]

1. The paper lacks important details about the data sources. The authors provide information about the web sources used for scraping Maithili data. However, the sources used for manually translated data are not mentioned.

A: Thank you for noting. We have corrected it and describe the specific source of the data are presented in Table 1, sec 2.4 and appendix A.1.

2. Similarly, the paper lacks details about the scraped/manually created data. Since the authors themselves decide the data sources, it would be beneficial to provide information about the domain of the sentence pairs (both scraped and manually translated). Also, details about length distribution etc. would have been beneficial.

A: The domain of the sentence pairs (both scraped and manually translated) is mentioned in Table 1, average sentence length we have shown in Table 5 and 6 of our dataset and existing NLLB dataset respectively.

3. The quality of the data is a major question. The authors scrape the data from the web and also use pyteserract for extracting text through OCR. Manual validation or cleaning of the extracted text doesn't seem to have been done. At least, there should have been a manual quality check about the sample of the dataset at each stage to ensure high data quality. Also, since the IndicTrans2 model is used for pseudo-parallel corpus generation from monolingual Maithili data, the quality of the pseudo-parallel corpus should have been validated through methods like COMET/SONAR/LaBSE/etc.

A: Thank you for your suggestions. The quality of manually created parallel data and pseudo-parallel data has been validated through LaBSE, LASER2, Median similarity, and standard deviation are shown in Table 2. Further details are provided in sec 2.4 on data quality.

4. There is no discussion about the absence of this data in existing open-source parallel corpora. There are parallel corpora like BPCC that already contain Hindi-Maithili parallel corpus. These datasets contain sentences available in the public domain, like on

the web. The authors do not carry any checks to ensure the newly generated data (pseudo or manual) is not a part of any existing available parallel corpora.

A: Thank you for noting this. BPCC parallel corpora has Maithili to English and vice-versa. and we have checked our newly generated data is not a part of any existing available parallel corpus like NLLB.

5. The creation of a test set is very naive, where the authors simply select random 10% sentence pairs and treat them as a test set. When existing Hindi-Maithili datasets are available (NLLB), I do not see any additional value in this test set. Rather, than creating a test set to cover a particular domain, creating a test set by handpicking certain types of sentence pairs (having linguistic variations, etc.) would be beneficial.

A: Thanks for your valuable suggestions. Our test set covers different domains such as story, novel, literature, news, culture, history types of sentence pairs, and we have compared the results of our dataset (MaitH 1.0) with the existing NLLB dataset shown in Table 4. The results on our dataset give better results than the existing dataset NLLB. It shows the importance of having manually curated and validated dataset as well as synthetic ones with longer sentences than present in NLLB.

6. The authors claim to give Hindi-Mathili MT baselines, which is not justfied given that there are already open-source MT models like NLLB, IndicTrans2, etc., which can translate between Hindi and Maithili.

A: Thanks for your valuable suggestions. The open-source MT models like NLLB, IndicTrans2 are available, but after finetuning all available models on our MaitH 1.0 dataset, they provide better results than the existing NLLB dataset. The results are shown in table 4. It shows the importance of a dataset which is clean, validated (partially), and longer sentences. Current SOTA for Maithili-Hindi is given by NLLB-200 which is outperformed by our dataset.

Reviewer [fwUM]

1. *The main weakness is lack of novelty in terms of methodology*

A: Pl see the response to Reviewer **[bz9c]**

2. *Comparative Evaluation Against IndicTrans2:*

A: We have compared the NLLB, mBART50 and mT5 model against the Indictrans2 model as shown in the Table 3 and Table 4.

Reviewer [y6JM]

1. Gathering and preparing parallel data for a very low-resource language pair, and then using it to train/fine-tune MT models, all of which is done using well-known methods, does not really provide any new or innovative contributions to the MT field, although the data itself may be useful in future research.

A: Thank you for your insightful feedback. In particular, we compared our manually curated dataset with the existing Maithili-Hindi parallel dataset available from NLLB sources. Our experiments show that training and testing baseline models (including IndicTrans2, mBART50, NLLB-200 and mT5) on the NLLB dataset resulted in significantly lower performance metrics like BLEU$, COMET, etc. compared to using our dataset. The results suggest that existing datasets, though large, for Maithili-Hindi might not be adequate for high-quality MT model training. This highlights the value of our dataset, which could serve as a benchmark for future research in the Maithili-Hindi language pair.

2. Typically, back translation is done from the high-resource language (HRL) to the low-resource language (LRL), so that when it is used to translate into the HRL (Hindi in this case), the model has at least learned how to produce fluent output in the HRL. I suspect that the results would have been better if this had been done.

A: Thanks for your suggestions. We have collected monolingual Maithili language and then generated the target (Hindi) sentences to increase the dataset and experiments show the benefits of synthetic dataset. In future we will try as you have suggested to generate synthetic data from HRL to LRL.

3. To show the effect of adding the back-translated data (which was likely of poor quality), you should have included results from fine-tuning mT5 and mBART only on the manually translated data.

A: We have finetuned the mT5, and mBART50 and also NLLB-200 models on our manually created Maithili to Hindi parallel dataset, results are shown in Table 3 in our paper.

4. You mention adhering to ethical principles including data privacy and consent in your Ethical Considerations section, but in section 3, you say you web-scraped your monolingual data but gave no indication of the licensing or consent associated with this data.

A: We have taken permissions over email.

5. There is no mention of making your parallel data available for others to use and to possibly reproduce your findings.

A: We have uploaded data and code on GitHub. Please refer to Appendix A.6 for code and reproducibility.