The Bangla/Bengali Seed Dataset Submission to the WMT24 Open Language Data Initiative Shared Task

Firoz Ahmed*
University of Florida
firozahmed@ufl.edu

Nitin Venkateswaran*
University of Florida
venkateswaran.n@ufl.edu

Sarah Moeller
University of Florida
smoeller@ufl.edu

Abstract

We contribute a seed dataset for the Bangla/Bengali language as part of the WMT24 Open Language Data Initiative shared task. We validate the quality of the dataset against a mined and automatically aligned dataset (NLLBv1) and two other existing datasets of crowdsourced manual translations. The validation is performed by investigating the performance of state-of-the-art translation models fine-tuned on the different datasets after controlling for training set size. Machine translation models fine-tuned on our dataset outperform models tuned on the other datasets in both translation directions (English-Bangla and Bangla-English). These results confirm the quality of our dataset. We hope our dataset will support machine translation for the Bangla/Bengali community and related low-resource languages.

1 Introduction

The Indian sub-continent is an area of rich linguistic diversity (Saxena and Borin, 2006; Hock and Bashir, 2016), and it is not uncommon for a language in this region to have both millions of speakers and insufficient resources for NLP development. Bangla/Bengali [ben] is one such language, ranked the 7th most spoken language in the world in the latest Ethnologue list of 200 most spoken languages (Eberhard and Fennig, 2024), and classified in the taxonomy of Joshi et al. (2020) as a Rising Star, "let down by insufficient efforts in labeled data collection" despite a "strong web presence and thriving online community". This classification contrasts squarely with that of languages such as Standard German, a "winner" in the Joshi et al. taxonomy because it has heavy investments in resources and technology, despite a ranking of 12 in the Ethnologue 200, below Bangla/Bengali.

The relative lack of data resources poses a challenge for neural machine translation (MT) efforts in Bangla/Bengali. While creating large-scale datasets of parallel sentences would be the next step towards improving Bangla/Bengali MT, efforts to create these types of resources have only recently been made (Hasan et al., 2020; Siripragada et al., 2020; Ramesh et al., 2022). Such efforts often must use automated methods to crawl and align the texts between language pairs, with manual checks and reviews being prohibitively expensive. There has been little work comparing larger datasets with smaller, professionally translated and manually curated datasets to investigate how the differences between these two types of dataset could impact the quality of machine translation.

This paper describes the results of one such manual effort, creating translation pairs between English and Bangla/Bengali for a smaller dataset and verifying the quality of those translations. Maillard et al. (2023) shows the sizeable impact of these smaller datasets on MT quality via bilingual and multilingual translation experiments, with the high quality manually translated datasets outperforming even a back-translation data augmentation approach with larger train set sizes. Continuing this line of reasoning, we hypothesize that models trained using a smaller but professionally translated dataset of Bangla/Bengali would perform better than models trained on larger, automatically mined and aligned parallel texts with little to no human intervention or review, once training sizes are controlled for. To this end, we created a smaller dataset of manual translations to test our hypotheses, and explored different training set sizes from larger datasets to check their equivalencies against our smaller dataset.

Our main contributions are as follows:

1. We contribute to the open datasets of the Open Language Data Initiative (OLDI) and produce a seed dataset for Bangla/Bengali by translat-

* These authors contributed equally

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 556–566
November 15–16, 2024.

ing the English seed dataset.

2. We carry out fine-tuning translation experiments to show that models tuned on our smaller, manually translated dataset outperform, or are on par with, models tuned on samples of comparable sizes from another dataset that has been automatically mined and aligned (NLLBv1).

3. We compare our dataset with other manually translated datasets for Bangla/Bengali available via OPUS (Tiedemann, 2012), and show that our dataset outperforms both corresponding and larger samples from these datasets (1.5x, 2x larger than our dataset) for a majority of pre-trained models in our experiments.

2 Related Work

Machine translation (MT) efforts in Bangla/Bengali currently rely on creative methods such as data augmentation and multilingual transfer to approach state-of-the-art MT. For example, Mondal et al. (2024), a recent work, uses back-translation to augment training data for English-Bengali transformer-based MT. Laskar et al. (2022a) augment data for English-Bengali MT using an SMT-based phrase-pair injection approach (Sen et al., 2021), and transliterate English texts into Bengali script as a transfer mechanism to share subword-level information between source and target sentences. Jasim et al. (2020) use a partial back-translation method by translating only selected phrases to the source language, achieving competitive results for Bengali MT on the WAT2018 (Nakazawa et al., 2018) test set. Laskar et al. (2022b) investigate knowledge transfer among Indic languages for neural MT, including Bengali, by transliterating all Indic languages into English script to share subword information during training. Bala Das et al. (2023) build a transformer-based multilingual neural MT system for 15 Indic language pairs, including Bengali, and English with shared encoder-decoders and transliteration schemes for related languages. Gala et al. (2023) build a multilingual NMT system for 22 Indic languages including Bengali.

Efforts to create large-scale datasets of parallel sentences for Bangla/Bengali have only recently been made. For example, Hasan et al. (2020) create a dataset of 2.75 million sentence pairs for machine translation, using an automated sentence segmentation toolkit and an ensemble of aligners for bitext alignment. Siripragada et al. (2020) collect parallel corpora across 10 Indian languages, including Bengali, by crawling two Government of India websites and applying document and sentence level alignment methods, producing 126.7K parallel texts for Bengali-English. Ramesh et al. (2022) create a dataset of parallel texts for 11 Indian languages, including 8.6 million parallel texts for Bengali-English, by crawling news and education/MOOC websites such as Coursera and passing the data through automated pipelines. Schwenk et al. (2021) mine billions of parallel texts from the web for multiple languages, which include approximately 10 million sentence pairs for Bengali-English aligned using LASER embeddings (Artetxe and Schwenk, 2019). As part of the No Language Left Behind project, NLLB Team et al. (2022) mine 62 million sentence pairs for Bengali-English using LASER3 embeddings (Heffernan et al., 2022); the NLLBv1 dataset created from the project is the largest known dataset of parallel texts for this language pair to date.

3 Meet The Data

Here we describe the language, the data collection process, and the format of the dataset.

3.1 Language description

Bangla/Bengali (ISO-639-3:ben, glottocode:beng1280, ISO-15924:Beng), an Indo-Aryan language, is the official and national language of Bangladesh and an official language of the state of West Bengal and other states in India. The language is commonly referred to as Bengali within the Indian states, and as Bangla in the nation of Bangladesh. The standardized dialects spoken in these two regions differ mainly in the morpho-phonological space. For example, Bangla has separate objective and genitive case markings for nouns and pronouns while Bengali has syncretized forms for these. The mid-back-rounded vowel (/ɔ/) is more common word-finally in Bangla than in Bengali. Despite "numerous small differences", both dialects have been called "indisputably the same language" (David, 2015). We refer to the language as Bangla for the rest of this paper, since the translations were produced in this dialect.

The script system of Bangla is similar to that of other South Asian languages in being an abugida

অন্ধকার	প্রায়	হলো	হলো।
ɔndhɔkar	*prae*	***ho-l-o***	***holo***
dark	almost	**become-PST-3.NHON**	**REDUP**

'It was just about **to get dark**. (Lit. The dark almost **happened happened**.)'

আমার	গাড়ি	পরিষ্কার	করতে	হবে।
amar	*gaṛi*	***porişkar***	***kor-te***	*hɔ-b-e*
1SG.GEN	car	**clean**	**do-IPFP**	become-FUT-3.NHON

'I need **to clean** my car.'

Figure 1: Glossed examples in Bangla script using reduplication and conjunct verbs; examples from (David, 2015)

system organized by syllables with two forms for each vowel viz., the independent and diacritic forms, and with a system of conjunct characters for complex consonant segments. The script (Fig. 1) is represented in Unicode with range 0980-09FF[1], which we use for our translations.

Bangla has certain features which make translation between Bangla and English a challenging task. These include rich morphological systems of inflection, derivation, and reduplication, a rich case system, a system of light verbs and conjunct verbs, and a system of noun classifiers. All these features are less prevalent in English. While Bangla is an SOV type language, scrambling of constituents within and across clauses for the purpose of altering information structure is common. This can pose a challenge for neural translation systems (Belinkov and Bisk, 2017). An in-depth description of the features of Bangla/Bengali can be found in David (2015).

3.2 Data Collection and Translation

The Bangla sentences in our dataset were manually translated from the English sentences in the Seed dataset v2.0 (Maillard et al., 2023) maintained by the Open Language Data Initiative. Details about the sourcing and composition of the dataset are described in Maillard et al. (2023). One native speaker of Bangla, an author of this paper fluent in English with graduate-level linguistic training and experience in professional translation from English to Bangla, translated all 6,193 sentences in the dataset. The Avro keyboard for Windows[2] with Unicode support was used to generate the Bangla translations. The translation guidelines[3] supplied by the Open Language Data Initiative were followed during the translation process.

3.3 Data Format

The Bangla translations are stored as a text file with a single line per translation, containing sentences in the same order as in the English seed dataset[4]. We follow the dataset formatting guidelines provided by the Open Language Data Initiative[5].

4 Experimental Validation

We compare our Bangla translations of the Seed dataset with the following three datasets. All datasets were downloaded from OPUS.

`NLLBv1` (NLLB Team et al., 2022). This is the largest available collection of automatically aligned Bangla-English sentence pairs with a wide range of text domains.

`Joshua-IPC` (Post et al., 2012). This is a dataset of parallel sentences for six Indic languages including Bangla. It was crowd-sourced by the authors via Amazon Mechanical Turk for translation experiments using the Joshua statistical MT system (Weese et al., 2011). Sentences for the Indic languages were extracted from the top 100 viewed Wikipedia pages for the language, and four English translations sourced for each sentence.

`TED2020` (Reimers and Gurevych, 2020). This is a dataset of crawled and aligned subtitles of TED Talks for the month of July 2020 across multiple languages, with subtitling carried out by a global community of volunteer translators[6]. We downloaded all 10,519 sentence pairs for Bangla-English, with translations in the English to Bangla direction.

4.1 Controlling for training set sizes

To facilitate comparisons with our translations, we control for training set sizes by sampling 1K, 3K and 6K sentence pairs from all datasets, similar to the approach used in Maillard et al. (2023). We select these training sizes to test whether models trained on smaller samples of our translations outperform models trained on samples of corresponding sizes from the other datasets. In addition, we sample 9K and 12K sentence pairs from the NLLBv1 and Joshua-IPC datasets, and 9K and the full 10,519 sentence pairs from TED2020. We

[1] https://www.unicode.org/charts/PDF/U0980.pdf
[2] https://www.omicronlab.com/avro-keyboard.html
[3] Translation guidelines:https://oldi.org/guidelines
[4] github:openlanguagedata/seed/blob/main/seed/eng_Latn
[5] Formatting guidelines: https://oldi.org/guidelines
[6] https://www.ted.com/participate/translate

compare results from these larger sizes with results trained on 6K sentence pairs from our translations.

We sample using five different seeds for each training size where possible, averaging results across all seeds instead of relying on results from a single sample per training size.

4.2 Translation models

We fine-tune translation models on existing pre-trained multilingual models and one pre-trained monolingual model in both directions (Bangla-English and English-Bangla). Only the sampled sentence pairs described in section 4.1 are used to fine-tune the models and no additional data is used. All models are fine-tuned using the HuggingFace `transformers` library (Wolf et al., 2020) and use a linear learning rate schedule with an initial rate of 1e-6, with warmup. The following pre-trained models are used.

`NLLB-200`. The state-of-the-art NLLB-200 model (NLLB Team et al., 2022) is pre-trained on 200 languages, including Bangla and English. The `nllb-200-1.3B` dense model with 1.3 billion parameters is used for our fine-tuning.

`mBART50` (Tang et al., 2020). This is a multilingual seq2seq model primarily intended for the task of machine translation through multilingual fine-tuning. This model is pre-trained on 50 languages, including Bangla and English. We use the `mbart-large-50` model.

`mT5` (Xue et al., 2021). We experiment with the multilingual variant of the text-to-text transformer pretrained on a Common Crawl based dataset containing 101 languages, including Bangla and English. The `mt5-large` model is used.

`BanglaT5` (Bhattacharjee et al., 2023). To investigate the impact of a pre-trained monolingual model on translation quality, we fine-tune the BanglaT5 model pretrained on the Bangla2B+ corpus (Bhattacharjee et al., 2022). We select this model based on its open-source availability and relatively large pre-training corpus size of 27.5GB. We note that future work could include experiments with other open-source pre-trained monolingual models as and when they become available.

4.3 Evaluation Metrics

We evaluate all models with the Bangla/Bengali and English datasets from the FLORES+ evaluation benchmark for multilingual machine translation (NLLB Team et al., 2022), maintained by the Open Language Data Initiative. We use the development set for model tuning and early stopping, and the test set to report translation metrics.

We report the chrF++ scores (Popović, 2017) calculated using the `sacrebleu` toolkit (Post, 2018), since the chrF-based score is known to correlate well with human rankings especially for morphologically rich languages like Bangla, outperforming BLEU (Popović, 2015). BLEU has also been shown to be less useful for morphologically complex languages, with language-specific customizations showing better correlations with human rankings (Chauhan et al., 2021; Bouamor et al., 2014).

5 Experiment Results

Comparing the results on our dataset against the others, we can confirm that our manual translations yielded high quality parallel sentences between English and Bangla. Tables 1 and 2 in the Appendix show the fine-tuned chrF++ scores in the English-Bangla and Bangla-English directions. Here we discuss the results in Figures 2, 3, 4 and 5, displayed below.

Displaying the results in the English to Bangla direction, Figure 2 shows that models fine-tuned on our translations outperform models tuned on samples of corresponding sizes from the other datasets. This demonstrates the high quality of our translations. In the case of NLLB-200 our translations are on par with the NLLBv1 samples. The results, interestingly, also hold for the 1K and 3K sample sizes showing that smaller samples of our translations are also effective. Given that it is likely the NLLBv1 dataset was used to pre-train the NLLB-200 models, it is not surprising that the NLLB-200 models fine-tuned on the NLLBv1 samples in our experiments show good performance despite the automatic sentence alignment process.

In Figure 3, it can be seen that in the Bangla to English direction, models tuned on our translations outperform other models across all corresponding sample sizes, except for NLLB-200 where models tuned on the TED2020 dataset show the best performance. The average performance gap for the 6K sample size between our translations and the other datasets is 5.34 points for the mBART50 model and 4.52 points for the mT5 model. These wide margins show that the quality of the dataset used for fine-tuning can make a sizeable difference even for multilingual models with large pre-training corpora in English.

Considering the test of our translations against

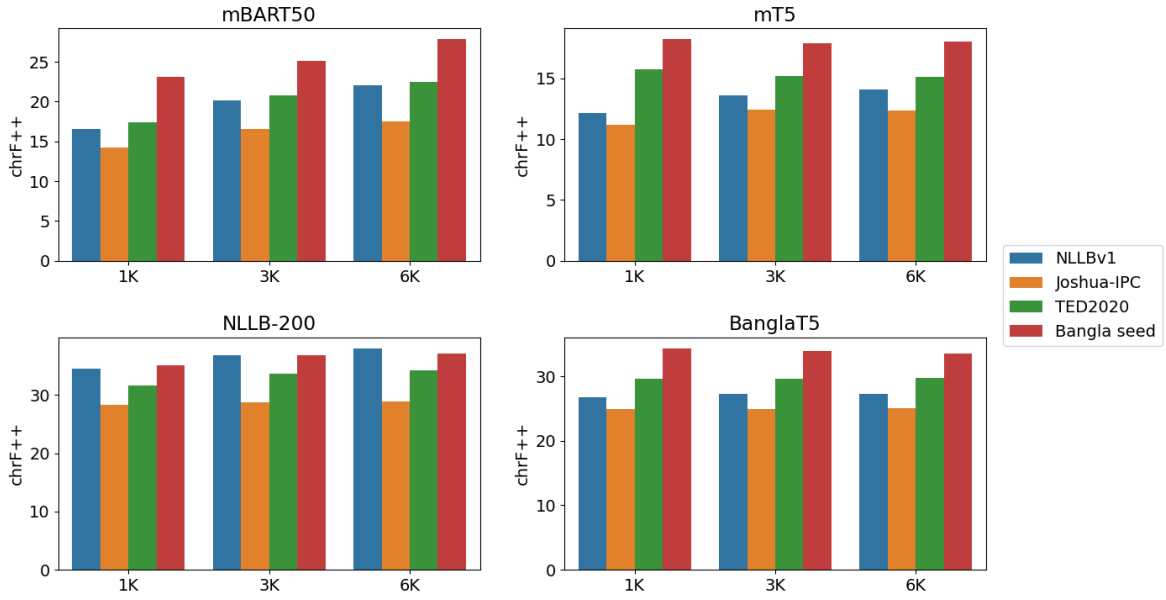

Figure 2: Averaged fine-tuned English to Bangla chrF++ scores on the FLORES+ test set for the 1K, 3K and 6K training set sizes. Models tuned on the Bangla seed dataset (red) outperform, or are on par with, models tuned on the other datasets across pre-trained model types and training sizes. Scores are averaged across five random samples per training set size and dataset.

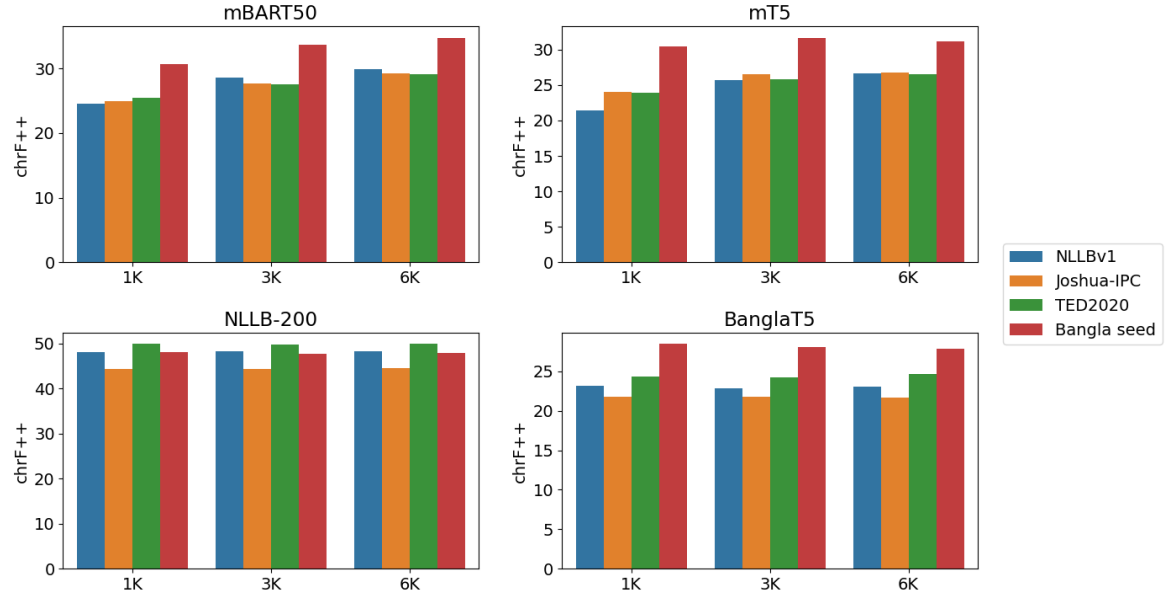

Figure 3: Averaged fine-tuned Bangla to English chrF++ scores on the FLORES+ test set for the 1K, 3K and 6K training set sizes. Models tuned on the Bangla seed dataset (red) outperform models tuned on the other datasets across pre-trained model types and training set sizes, except for the NLLB-200 models. Scores are averaged across five random samples per training set size and dataset.

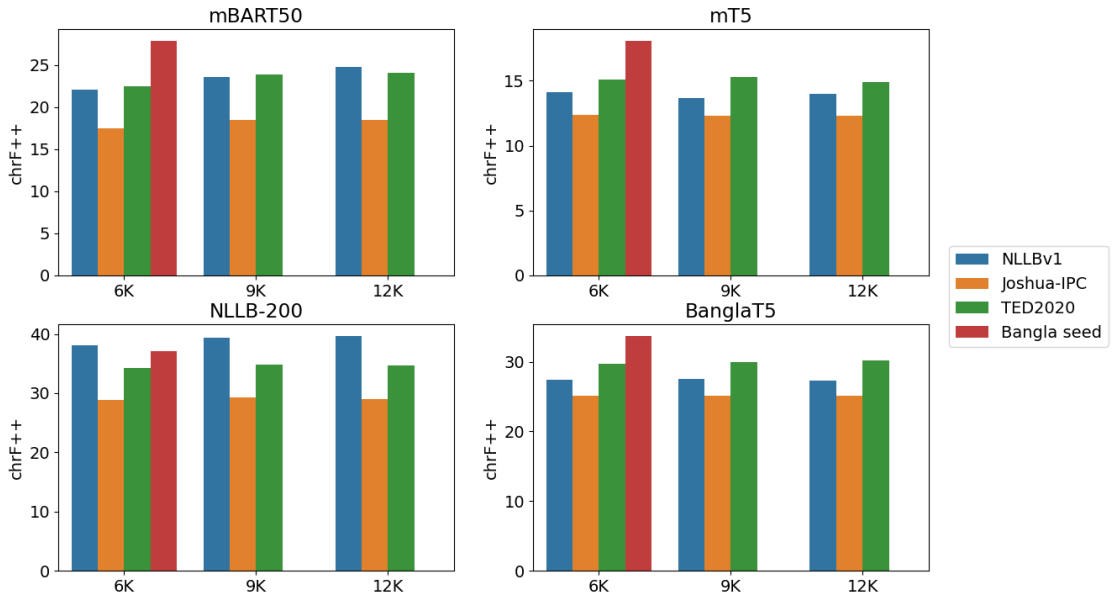

Figure 4: Averaged fine-tuned English to Bangla chrF++ scores on the FLORES+ test set for the 6K, 9K and 12K training set sizes; the complete TED2020 dataset is used in the 12K case. Models tuned on the Bangla seed dataset (red) outperform models tuned on other datasets of larger training sizes (9K, 12K) across pre-trained model types, except for the NLLB-200 models tuned on the NLLBv1 data. Scores are averaged across five random samples per training set size and dataset.

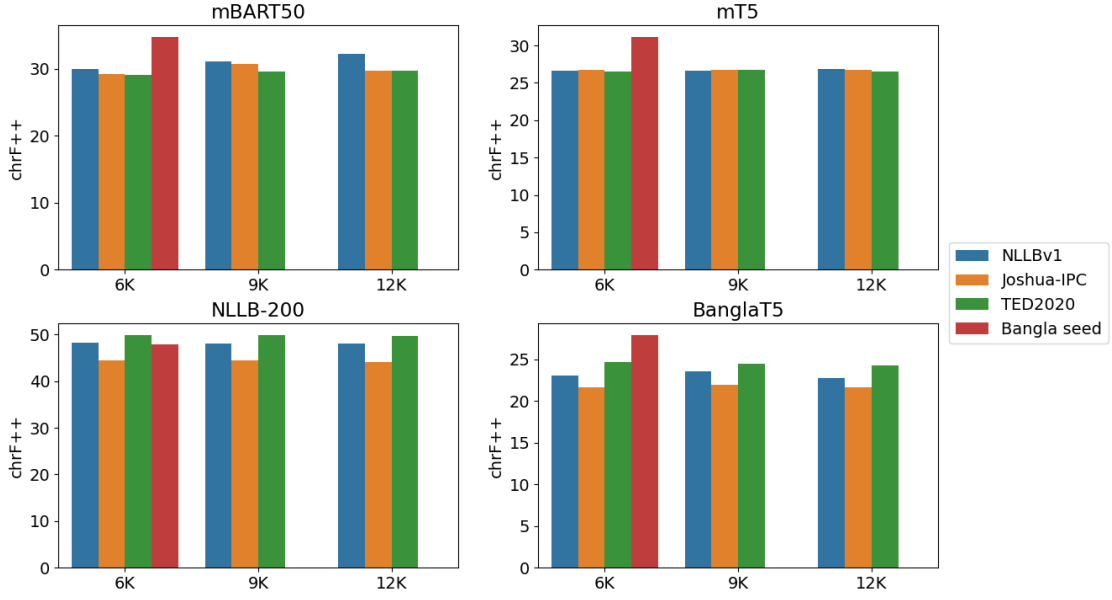

Figure 5: Averaged fine-tuned Bangla to English chrF++ scores on the FLORES+ test set for the 6K, 9K and 12K training set sizes; the complete TED2020 dataset is used in the 12K case. Models tuned on the Bangla seed dataset (red) outperform models tuned on other datasets of larger training sizes (9K, 12K) across pre-trained model types, except for the NLLB-200 models. Scores are averaged across five random samples per training set size and dataset.

samples of larger sizes from the other datasets, i.e 9K and 12K sentence pairs. The results in Figures 4 and 5 show that our translations outperform these larger training samples across all pre-trained model types and datasets except for the NLLB-200 models. The NLLB-200 models tuned on larger samples of the NLLBv1 dataset in the English to Bangla direction scored better than our translations. This is as expected, given the possible overlap between the NLLBv1 datasets used to pre-train and fine-tune the models in our experiments. The fact that models tuned on our translations scored better than models tuned on larger samples from the other datasets is another demonstration of the higher quality of our dataset.

6 Conclusion

We have created a high quality dataset of Bangla-English seed translations to contribute to the Open Language Data Initiative, paving the way for more translations between Bangla and other languages, including low-resource ones, that are supported by the initiative. We have demonstrated the high quality of our translated dataset by comparing it with a larger dataset that was mined and automatically aligned, as well as with two datasets of crowdsourced and reviewed translations. The models tuned on our dataset outperform models tuned on the other datasets after controlling for training set size. We hope that our dataset will support ongoing research in machine translation for the Bangla/Bengali community and other low-resource languages.

Acknowledgements

We would like to thank the organizers of the Open Language Data Initiative Shared Task for their support throughout the process.

Ethics Statement

Licensing We release the dataset under a CC-BY-SA-4.0 license.

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.

Sudhansu Bala Das, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kr. Patra. 2023. Improving multilingual neural machine translation system for indic languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6).

Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *ArXiv*, abs/1711.02173.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. 2022. BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1318–1327, Seattle, United States. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2023. BanglaNLG and BanglaT5: Benchmarks and resources for evaluating low-resource natural language generation in Bangla. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 726–735, Dubrovnik, Croatia. Association for Computational Linguistics.

Houda Bouamor, Hanan Alshikhabobakr, Behrang Mohit, and Kemal Oflazer. 2014. A human judgement corpus and a metric for Arabic MT evaluation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 207–213, Doha, Qatar. Association for Computational Linguistics.

Shweta Singh Chauhan, Philemon Daniel, Archita Mishra, and Abhay Kumar. 2021. Adableu: A modified bleu score for morphologically rich languages. *IETE Journal of Research*, 69:5112 – 5123.

Anne Boyle David. 2015. *Descriptive Grammar of Bangla*. De Gruyter Mouton, Berlin, München, Boston.

Gary F. Simons Eberhard, David M. and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, twenty-seventh edition. SIL International, Dallas, TX, USA.

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*.

Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 2020. Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

pages 2612–2623, Online. Association for Computational Linguistics.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hans Henrich Hock and Elena Bashir, editors. 2016. *The Languages and Linguistics of South Asia*. De Gruyter Mouton, Berlin, Boston.

Binu Jasim, Vinay Namboodiri, and C V Jawahar. 2020. PhraseOut: A code mixed data augmentation method for MultilingualNeural machine tranlsation. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 470–474, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Sahinur Rahman Laskar, Pankaj Dadure, Riyanka Manna, Partha Pakray, and Sivaji Bandyopadhyay. 2022a. English to Bengali multimodal neural machine translation using transliteration-based phrase pairs augmentation. In *Proceedings of the 9th Workshop on Asian Translation*, pages 111–116, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Sahinur Rahman Laskar, Riyanka Manna, Partha Pakray, and Sivaji Bandyopadhyay. 2022b. Investigation of multilingual neural machine translation for indian languages. In *Workshop on Asian Translation*.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzmán. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Subrota Kumar Mondal, Chengwei Wang, Yijun Chen, Yuning Cheng, Yanbo Huang, Hong-Ning Dai, and H. M. Dipu Kabir. 2024. Enhancement of english-bengali machine translation leveraging back-translation. *Applied Sciences*, 14(15).

Toshiaki Nakazawa, Katsuhito Sudoh, Shohei Higashiyama, Chenchen Ding, Raj Dabre, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, and Sadao Kurohashi. 2018. Overview of the 5th workshop on Asian translation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation: 5th Workshop on Asian Translation*, Hong Kong. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six Indian languages via crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada. Association for Computational Linguistics.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Kumar Deepak, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 Indic languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Anju Saxena and Lars Borin, editors. 2006. *Lesser-Known Languages of South Asia*. De Gruyter Mouton, Berlin, New York.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Sukanta Sen, Mohammed Hasanuzzaman, Asif Ekbal, Pushpak Bhattacharyya, and Andy Way. 2021. Neural machine translation of low-resource languages using smt phrase pair injection. *Natural Language Engineering*, 27(3):271–292.

Shashank Siripragada, Jerin Philip, Vinay P. Namboodiri, and C V Jawahar. 2020. A multilingual parallel corpora collection effort for Indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3743–3751, Marseille, France. European Language Resources Association.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401.

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).

Jonathan Weese, Juri Ganitkevitch, Chris Callison-Burch, Matt Post, and Adam Lopez. 2011. Joshua 3.0: Syntax-based machine translation with the thrax grammar extractor. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 478–484, Edinburgh, Scotland. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

A Appendix

A.1 Fine-tuned English to Bangla chrF++ scores

Table 1 shows the fine-tuned English to Bangla chrF++ scores across all model types, datasets, and training set sizes.

A.2 Fine-tuned Bangla to English chrF++ scores

Table 2 shows the fine-tuned Bangla to English chrF++ scores across all model types, datasets, and training set sizes.

A.3 `sacrebleu` version string

The `sacrebleu` version string is provided below for reproducibility: `nrefs:1|case:mixed|eff:yes|nc:6|nw:2|space:no|version:2.4.2`

	mBART50				
	1K	3K	6K	9K	12K
NLLBv1	16.6	20.12	22.04	23.6	24.72
Joshua-IPC	14.22	16.58	17.48	18.42	18.44
TED2020	17.42	20.8	22.48	23.82	24.04
Bangla seed	**23.12**	**25.14**	**27.82**	–	–
	mT5				
	1K	3K	6K	9K	12K
NLLBv1	12.12	13.62	14.12	13.66	14
Joshua-IPC	11.2	12.46	12.34	12.32	12.28
TED2020	15.74	15.22	15.1	15.28	14.9
Bangla seed	**18.2**	**17.88**	**18.04**	–	–
	NLLB-200				
	1K	3K	6K	9K	12K
NLLBv1	34.5	36.84	**38.04**	**39.3**	**39.6**
Joshua-IPC	28.32	28.76	28.92	29.3	29
TED2020	31.6	33.7	34.26	34.76	34.64
Bangla seed	**35.12**	**36.9**	37.14	–	–
	BanglaT5				
	1K	3K	6K	9K	12K
NLLBv1	26.76	27.36	27.36	27.52	27.32
Joshua-IPC	24.98	25	25.06	25.14	25.08
TED2020	29.64	29.7	29.76	29.94	30.2
Bangla seed	**34.4**	**33.92**	**33.64**	–	–

Table 1: Average fine-tuned English to Bangla chrF++ scores on the FLORES+ test set. Scores are averaged across five random samples per training set size and dataset

	mBART50				
	1K	3K	6K	9K	12K
NLLBv1	24.54	28.62	29.94	31.14	32.2
Joshua-IPC	24.96	27.68	29.18	30.08	29.76
TED2020	25.46	27.5	29.12	29.58	29.54
Bangla seed	**30.62**	**33.64**	**34.76**	–	–
	mT5				
	1K	3K	6K	9K	12K
NLLBv1	21.42	25.68	26.6	26.66	26.84
Joshua-IPC	24.06	26.46	26.74	26.74	26.78
TED2020	23.88	25.84	26.46	26.76	26.48
Bangla seed	**30.46**	**31.66**	**31.12**	–	–
	NLLB-200				
	1K	3K	6K	9K	12K
NLLBv1	48.16	48.22	48.3	48.08	48.14
Joshua-IPC	44.36	44.42	44.54	44.48	44.14
TED2020	**49.88**	**49.74**	**49.94**	**49.88**	**49.7**
Bangla seed	48	47.78	47.84	–	–
	BanglaT5				
	1K	3K	6K	9K	12K
NLLBv1	23.2	22.5	23.08	23.4	23.2
Joshua-IPC	21.74	21.82	21.66	21.98	21.68
TED2020	24.38	24.28	24.66	24.44	24.3
Bangla seed	**28.48**	**28.1**	**27.92**	–	–

Table 2: Average fine-tuned Bangla to English chrF++ scores on the FLORES+ test set. Scores are averaged across five random samples per training set size and dataset