First Attempt at Building Parallel Corpora for Machine Translation of Northeast India's Very Low-Resource Languages

Atnafu Lambebo Tonja[1,2,3], **Melkamu Mersha** [1], **Ananya Kalita**[4],
Olga Kolesnikova [2] **and Jugal Kalita** [1]

[1] University of Colorado, Colorado Springs, USA [2] Instituto Politécnico Nacional, Mexico city, Mexico,
[3] Lelapa AI, South Africa, [4] Palmer Ridge High School, Monument, Colorado, USA

Abstract

This paper presents the creation of initial bilingual corpora for thirteen very low-resource languages of India, all from Northeast India. It also presents the results of initial translation efforts in these languages. It creates the first-ever parallel corpora for these languages and provides initial benchmark neural machine translation results for these languages. We intend to extend these corpora to include a large number of low-resource Indian languages and integrate the effort with our prior work with African and American-Indian languages to create corpora covering a large number of languages from across the world.

1 Introduction

In the last few years, there have been significant advancements in deep learning approaches, such as the development of transformers (Vaswani et al., 2017). These advancements have greatly improved machine translation (MT), a core natural language processing (NLP) task. Regarding coverage and translation quality, MT has shown remarkable improvements (Wang et al., 2021). Most models and methods for high-resource languages do not perform well in low-resource settings. Low-resource languages have also suffered from inadequate language technology designs (Costa-jussà et al., 2022; Tonja et al., 2023a; King, 2015; Joshi et al., 2019; Tonja et al., 2022, 2023b; Yigezu et al., 2021). Creating effective methods for natural language tasks is challenging, with extremely limited resources and little to no available data. The problem becomes worse without a parallel dataset for a vast number of world's languages (Joshi et al., 2020; Ranathunga et al., 2023; Adebara and Abdul-Mageed, 2022).

Ethnologue[1] enumerates 7,618 extant human languages in the world, of which it estimates 3,072 are endangered. India has 30 languages spoken by more than a million native speakers, 92 additional languages spoken by more than 10,000 people, and 1599 other languages. The terms language and dialect may have been conflated in Census of India reports[2]. Of these, 197 languages are endangered (Dash et al., 2022). The languages with which work in this paper are a small fraction of such languages.

This paper makes a selection of low-resource languages of India, all from Northeast India, with the intention of developing resources to facilitate neural machine translation. We choose Northeast India because this area has a very high concentration of very low-resource and/or endangered languages, and the area being remote from mainland India has exacerbated scholarly neglect.

2 Related Work

Dash (2020) reviewed the issues languages with a small number of speakers face in India. Based on extensive research, the author surmised that the amount of published work on low-resource and endangered languages of India, especially in the context of technology and, in particular, natural language processing, is miniscule.

Chauhan et al. (2021) provided a monolingual corpus of Kangri (a language spoken by 1.1 million people in Himachal Pradesh and Punjab) with 1.8M words, as well bitext Hindi-Kangri corpus of 27 thousand words. For NMT, they used a model that learns to map word embedding from one language to another in an unsupervised manner (Artetxe et al., 2018). Only cursory details are given regarding the architecture of the neural model, although translation metrics are given in terms of BLEU and Meteor scores. The corpora are available on Github[3].

Acharya (2021) discussed a proposal for building a digital archive to collect and preserve textual,

[1] www.ethnologue.com

[2] https://en.wikipedia.org/wiki/Languages_of_India

[3] https://github.com/chauhanshweta/Kangri_corpus

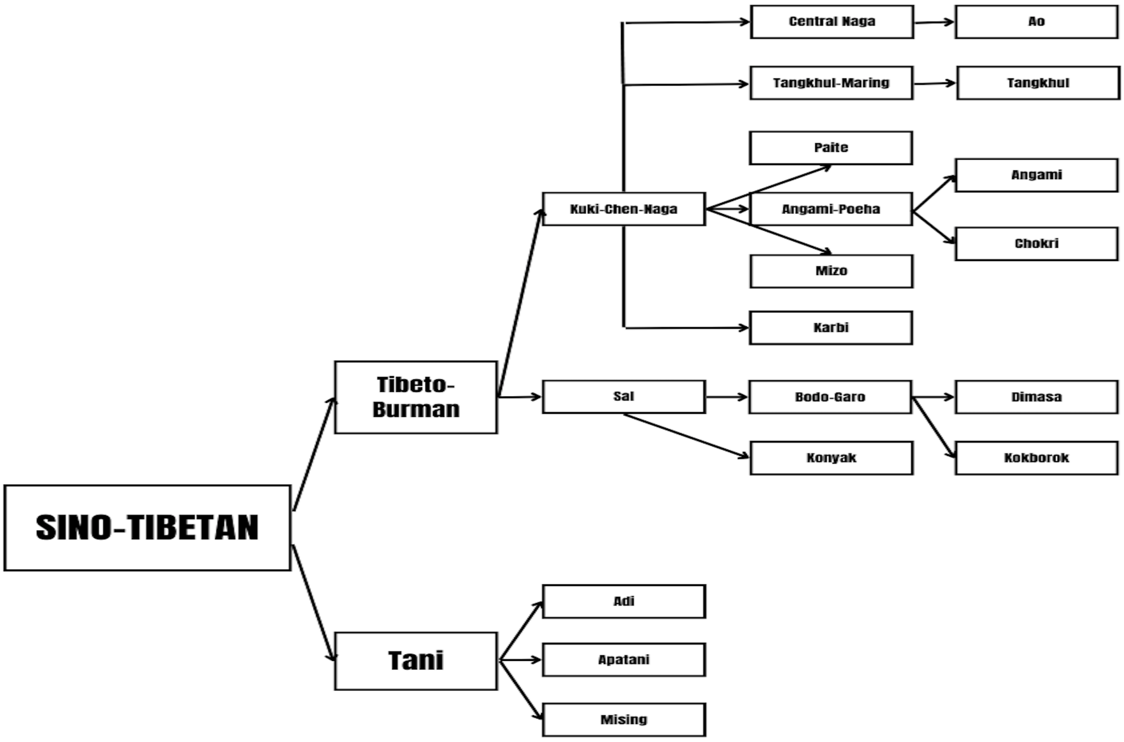

Figure 1: The languages under consideration fall under the Sino-Tibetan family and two subfamilies, Tani and Tibeto-Burman.

audio, and video documentation of twelve Munda languages, a sub-family of Austro-Asiatic languages. Of twenty Munda languages, spoken by between just 20 to 7M people, UNESCO has identified twelve as endangered. The authors proposed to use advanced technologies like artificial intelligence in the design of the archive. However, it is not clear what has been achieved.

To the authors' knowledge, there is no prior computational work of any kinda in the very low-resource and endangered languages of Northeast India.

3 Languages Under Consideration

The Sino-Tibetan language family includes more than 400 modern languages spoken in China, India, Burma, and Nepal. It is one of the most diverse language families in the world, with 1.4 billion speakers, including Chinese, Tibetan, and Burmese. Based on a phylogenetic study of 50 ancient and modern Sino-Tibetan languages, scholars have recently concluded that the Sino-Tibetan languages originated in North China around 7,200 years ago (Sagart et al., 2019). Various classification schemes have been proposed for the Sino-Tibetan family of languages (Matisoff, 2003, 2015; Driem, 2001). Our discussion follows Matisoff's classification that divides Sino-Tibetan languages into a number of sub-families at various levels.

Table 1 provides a list of languages we work with in this paper. No detailed description of the languages is given in this paper due to lack of space. The languages are from Northeast India, which is linguistically very diverse and has a large number of very low-resource languages that are vulnerable or endangered. The languages we have chosen all belong to the Sino-Tibetan family. Under this class, there are two sub-families called Tibeto-Burman and Taani, among other sub-families. All our languages come from these two sub-families (see Figure 1).

4 Dataset

4.1 Dataset Collection

We obtained datasets in 13 Indian languages from religious domains through a Bible-related website[4]. To extract the Bible data from these websites, we utilized a web crawler that identified the structure of web documents, including pages, books, and phrases, for each article. Python libraries like *requests*, regular expressions (*R*), and Beautiful Soup (*BS*) were used to extract article content and ana-

[4] https://www.bible.com/

Language	ISO Code	Family	Speakers	Location	Corpus Domain	Corpus Size
Adi	adi	Tani	150K	Arunachal Pradesh	Religious	29301
Angami	njm	Tibeto-Burman	150K	Nagaland	"	30017
Ao	njo	Tibeto-Burman	607K	Nagaland	"	29121
Apatani	apt	Tani	45K	Arunachal Pradesh	"	7185
Chokri	nri	Tibeto-Burman	111K	Nagaland	"	7821
Dimasa	dis	Tibeto-Burman	137K	Assam, Nagaland	"	10275
Karbi	mjw	Tibeto-Burman	2.5M	Assam, Meghalaya, Arunachai Pradesh	"	7185
Kokborok	trp	Tibeto-Burman	1M	Tripura, Assam, Mizoram, Myanmar, Bangladesh	"	29298
Konyak	nbe	Tibeto-Burman	246K	Nagaland, Myanmar	"	28518
Mising	mrg	Tani	629K	Assam	"	7825
Paite	pck	Tibeto-Burman	1M	Manipur, Mizoram, Assam, Myanmar	"	29615
Tangkhul	nmf	Tibeto-Burman	140K	Manipur, Nagaland	"	28324
Thado	tcz	Tibeto-Burman	350K	Manipur, Nagaland, Assam, Mizoram	"	29004

Table 1: Languages Under Consideration. Accurate population count is difficult to obtain and varies substantially among sources, corpus size shows the number of parallel sentences.

lyze website structure from a given URL. Extensive research did not discover any additional publicly available texts in these languages.

4.2 Sentence Alignment

We gathered the limited corpora for various languages and aligned each Indian language sentence with a corresponding sentence in English to create a dataset for the MT experiment. We followed the heuristic alignment method outlined by the Tonja et al. (2023c) to align the sentences.

4.3 Dataset Pre-processing

We aligned the texts of Indian languages with their corresponding translations in English. Before splitting the corpus, we pre-processed by removing numbers, special characters, and sentences that contain less than five words. We divided the pre-processed corpus into training, development, and test sets in a 70:10:20 ratio for the baseline experiments. Detailed information on the selected languages, language families, domain, and dataset size can be found in Table 1.

5 Baseline Experiments and Results

5.1 Experiments

To evaluate the usability of the newly collected corpus, we trained bi-directional MT models that can translate Indian languages to/from English using (1) **transformer** and (2) **fine-tuning** multilingual machine translation model.

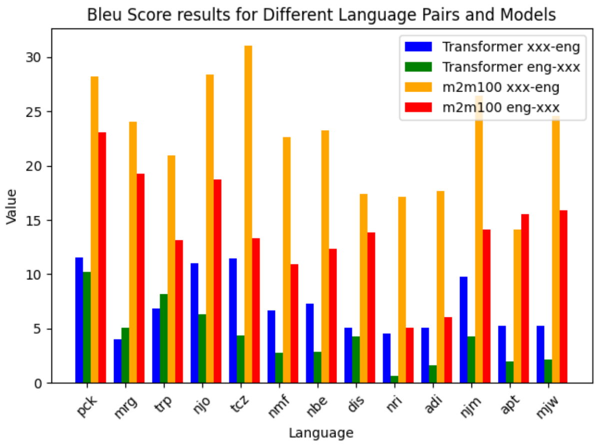

Figure 2: Benchmark translation results for transformer and fine-tuned approaches in both directions (English/low-resource Indian languages).

1) Transformer - is a type of neural network architecture first introduced in the paper *Attention Is All You Need* (Vaswani et al., 2017). Transformers are state-of-the-art approaches widely used in NLP tasks such as MT, text summarization, and sentiment analysis. We trained transformers from scratch for this experiment.

2) Fine-tuning involves using a pre-trained MT model and adapting it to a specific translation task, such as translating between a particular language pair or in a specific domain (Lakew et al., 2019). We used **M2M100-48** a multilingual encoder-decoder (seq-to-seq) model trained for many-to-many multilingual translation (Fan et al., 2020). We used a model with 48M parame-

Model	en-xx													
	adi	**apt**	**dis**	**mjw**	**mrg**	**nbe**	**njm**	**njo**	**nmf**	**nri**	**pck**	**tcz**	**trp**	**Avg.**
	Bleu Score													
Transformer	1.62	2	4.33	2.17	5.12	2.91	4.26	6.32	2.81	0.66	10.23	4.35	9.18	4.30
m2m100-fine-tuned	6.06	15.49	13.82	15.91	19.23	12.31	14.07	18.76	10.88	5.10	23.03	13.28	13.16	**13.93**

Table 2: Benchmark translation results from English to low-resource Indian languages

Model	xx-en													
	adi	**apt**	**dis**	**mjw**	**mrg**	**nbe**	**njm**	**njo**	**nmf**	**nri**	**pck**	**tcz**	**trp**	**Avg.**
	Bleu Score													
Transformer	5.09	5.29	5.11	5.3	4.02	7.32	9.8	11.06	6.64	4.51	11.54	11.49	6.88	7.63
m2m100-fine-tuned	17.62	14.11	17.42	24.52	24.03	23.27	26.44	28.37	22.61	17.12	28.16	31.03	20.90	**22.67**

Table 3: Benchmark translation results from low-resource Indian Languages to English language

ters due to computing resource limitations.

5.2 Results

We used Sacrebleu (Post, 2018) evaluation metrics to evaluate translation models. Tables 2, 3 and Figure 2 show the translation results in both directions (to/from English - from/to low-resource Indian languages)

5.2.1 Translating from English to low-resource Indian languages

In Table 2, we present the translation results from English to low-resource Indian languages. We observed that fine-tuning the m2m100 model performs better than using a transformer trained from scratch. The transformer model's performance also varies significantly (0.66 – 10.23 spBLEU) depending on the language and corpus size. This indicates that a bilingual translation model trained from scratch performs poorly for low-resource language training compared to fine-tuning multilingual translation models. Fine-tuning the multilingual model produced better results than the model built from scratch for English to Indian language translation.

5.2.2 Translating from low-resource Indian languages to English

Table 3 displays the results of using English as the target language to translate low-resource Indian languages. As is evident from the results, the fine-tuned model outperforms the transformer model significantly when translating from Indian languages to English. However, when it comes to translating similar languages to English, the transformer model shows an improvement compared to Table 2. It is worth noting that the fine-tuned model exhibits better Bleu scores while translating to English than when translating to low-resource Indian languages. The results indicate that languages with larger datasets tend to perform better. Therefore, both models exhibit improved performance while translating from low-resource Indian languages to English, whereas the model struggles to translate from English to low-resource Indian languages.

6 Conclusions and Future Work

This paper presents the first, albeit limited size, parallel corpus for 13 low-resource Sino-Tibetan Indian languages paired with English and discusses the benchmark results for the translation of language pairs to/from low-resource Indian languages from/to English. We evaluated the usability of the collected corpus by using transformer and fine-tuning multilingual translation model. From our results fine-tuning multilingual model outperforms transformer model trained from scratch in both translation directions.

In the future, we aim to increase corpus sizes of these low-resource languages by extracting text from scanned documents if/where available and evaluate additional machine translation approaches to improve performance. We intend to increase the number of languages substantially by first incorporating all low-resource languages of India for which a Bible translation exists. We also plan to find language communities in social media platforms such as Facebook and attempt to gather additional bitext documents and evaluate the quality of translations with native speakers.

References

Kaushik Acharya. 2021. KaushikAcharya at SemEval-2021 task 9: Candidate generation for fact verification over tables. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1271–1275, Online. Association for Computational Linguistics.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. *arXiv preprint arXiv:1805.06297*.

Shweta Chauhan, Shefali Saxena, and Philemon Daniel. 2021. Monolingual and parallel corpora for kangri low resource language. *arXiv preprint arXiv:2103.11596*.

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*.

Rajendra Kumar Dash. 2020. Revitalizing endangered languages in india: Can public-private partnership (ppp) work. In *2nd International Conference on Social Sciences in the 21st Century, Seminar Paper*.

Sarthak Dash, Sugato Bagchi, Nandana Mihindukulasooriya, and Alfio Gliozzo. 2022. Permutation invariant strategy using transformer encoders for table understanding. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 788–800, Seattle, United States. Association for Computational Linguistics.

George van Driem. 2001. Languages of the himalayas: an ethnolinguistic handbook of the greater himalayan region: containing an introduction to the symbiotic theory of language. *(No Title)*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *arXiv preprint*.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. *arXiv preprint arXiv:1912.03457*.

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. *arXiv preprint arXiv:2004.09095*.

Benjamin Philip King. 2015. *Practical Natural Language Processing for Low-Resource Languages*. Ph.D. thesis.

Surafel M Lakew, Alina Karakanta, Marcello Federico, Matteo Negri, and Marco Turchi. 2019. Adapting multilingual neural machine translation to unseen languages. *arXiv preprint arXiv:1910.13998*.

James A Matisoff. 2003. *Handbook of Proto-Tibeto-Burman: system and philosophy of Sino-Tibetan reconstruction*. Univ of California Press.

James A Matisoff. 2015. *The Sino-Tibetan etymological dictionary and thesaurus*. Regents of the University of California.

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, Brussels, Belgium. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin J Ryder, Valentin Thouzeau, Simon J Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of sino-tibetan. *Proceedings of the National Academy of Sciences*, 116(21):10317–10322.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. 2023a. Natural language processing in ethiopian languages: Current state, challenges, and opportunities. *arXiv preprint arXiv:2303.14406*.

Atnafu Lambebo Tonja, Olga Kolesnikova, Muhammad Arif, Alexander Gelbukh, and Grigori Sidorov. 2022. Improving neural machine translation for low resource languages using mixed training: The case of ethiopian languages. In *Mexican International Conference on Artificial Intelligence*, pages 30–40. Springer.

Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh, and Grigori Sidorov. 2023b. Low-resource neural machine translation improvement using source-side monolingual data. *Applied Sciences*, 13(2):1201.

Atnafu Lambebo Tonja, Christian Maldonado-Sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023c. Parallel corpus for indigenous language translation: Spanish-mazatec and spanish-mixtec. *arXiv preprint arXiv:2305.17404*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2021. Progress in machine translation. *Engineering*.

Mesay Gemeda Yigezu, Michael Melese Woldeyohannis, and Atnafu Lambebo Tonja. 2021. Multilingual neural machine translation for low resourced languages: Ometo-english. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 89–94. IEEE.