Towards Resource-Rich Mizo and Khasi in NLP: Resource Development, Synthetic Data Generation and Model Building

Soumyadip Ghosh
IIIT Hyderabad
soumya50052@gmail.com

Henry Lalsiam
North-Eastern Hill University
neihsialhenry25@gmail.com

Dorothy Marbaniang
Assam University, Silchar
dolly.marbz@gmail.com

Gracious Mary Temsen
University of Hyderabad
gmtemsen@uohyd.ac.in

Rahul Mishra
IIIT Hyderabad
rahul.mishra@iiit.ac.in

Parameswari Krishnamurthy
IIIT Hyderabad
param.krishna@iiit.ac.in

Abstract

In the rapidly evolving field of Natural Language Processing (NLP), Indian regional languages remain significantly underrepresented due to their limited digital presence and lack of annotated resources. This work presents the first comprehensive effort toward developing high quality linguistic datasets for two extremely low resource languages Mizo and Khasi. We introduce human annotated, gold standard datasets for three core NLP tasks: Part-of-Speech (POS) tagging, Named Entity Recognition (NER), and Keyword Identification. To overcome annotation bottlenecks in NER, we further explore a synthetic data generation pipeline involving translation from Hindi and cross-lingual word alignment. For POS tagging, we adopt and subsequently modify the Universal Dependencies (UD) framework to better suit the linguistic characteristics of Mizo and Khasi, while custom annotation guidelines are developed for NER and Keyword Identification. The constructed datasets are evaluated using multilingual language models, demonstrating that structured resource development, coupled with gradual fine-tuning, yields significant improvements in performance. This work represents a critical step toward advancing linguistic resources and computational tools for Mizo and Khasi.

1 Introduction

India is home to more than 1,963 languages (Census Commissioner, 2022), belonging to five major language families, yet the Indian Constitution officially recognizes only 22 (Indian-Constitution, 2022). While English and Hindi are spoken by approximately 10.2% and 43.63% of the population, respectively, the majority prefer using their regional languages. However, a vast number of these languages remain underrepresented in the field of Natural Language Processing (NLP), primarily due to the lack of curated resources and limited availability of digital text in native scripts. While high-resource languages benefit from abundant datasets, extremely low-resource languages like Mizo and Khasi (Sarkar et al., 2024) have very limited digital presence.

Example 1:

English: The sun is shining in the sky.

Khasi : Ka sngi ka shai thaba ha ka suin bneng.

Mizo: Ni chu vânah a ên mêk .

Example 2:

English: While thanking the Garo people on this day, the Registrar General High Court of Meghalaya, Mr E. Kharumnuid expressed his praise to the members of the Wangala Committee that he gets to witness this Festival.

Khasi: Haba ai ka jingkhublei sha ki jaitbynriew Garo ha kane ka sngi u Registrar General, High Court ka Meghalaya, u Bah E Kharumniud u la pynpaw ka jingïaroh ïa ki dkhot ka Committee jong ka Wangala kaba u la ïoh ban sakhi ïa kane ka tamasa.

Mizo: Hemi ni hian Garo mipuite chungah lawmthu a sawi rualin, Registrar General High Court of Meghalaya, Mr E. Kharumnuid chuan he Festival hi a hmuh theih avangin Wangala Committee member-te chu a fak thu a sawi.

Figure 1: Example sentences in Mizo and Khasi with their corresponding English translations.

Mizo, a Tibeto-Burman language (Thurgood and LaPolla, 2003), is spoken by approximately 831K people, while Khasi, an Austroasiatic language (Jenny and Sidwell, 2014), is spoken by around 1.4M (according to Census 2011) people in India. A more comprehensive linguistic description of Mizo can be found in Appendix A.1, and for Khasi in Appendix A.2. Figure 1 illustrates example sentences in Mizo and Khasi corresponding to the same English sentence.

In this work, we focus on the development of foundational linguistic resources to support NLP for Mizo and Khasi. Specifically, we created datasets for Part-of-Speech (POS) (Kumar et al., 2024) tagging, Named Entity Recognition (NER) (Murthy et al., 2018) and Keyword Identification (Bala et al., 2024).Given the lack of task-specific

annotation guidelines for Mizo and Khasi, we adapted the Universal Dependencies (UD) (Universal Dependencies, 2025) framework for POS tagging and designed custom annotation schemes that reflect the unique syntactic and semantic characteristics of these two languages. Additionally, we created separate annotation guidelines for NER and Keyword Identification to ensure accurate dataset construction for each task.

To mitigate the challenge posed by the scarcity of gold-standard annotated data, especially for NER, we explored synthetic data generation using a Hindi NER dataset as a source. This involved translation into Mizo and Khasi, followed by word alignment using models such as Awesome-Align (Dou and Neubig, 2021) and VecMAP (Artetxe et al., 2017, 2018). The alignment process was carefully evaluated and refined to ensure the quality and usability of the resulting synthetic annotations. However, existing language models exhibit little to no understanding of Mizo and Khasi. To bridge this gap, we first constructed a monolingual corpus for both languages and performed multistage fine-tuning of multilingual models such as MuRIL (Khanuja et al., 2021), RemBERT (Conneau et al., 2019), and XLM-RoBERTa-Large (Chung et al., 2021).

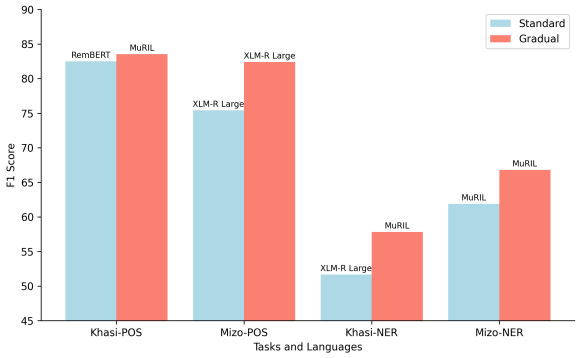

Figure 2: Comparison of best-performing models under standard and gradual fine-tuning approaches across different tasks in Khasi and Mizo. The best-performing models for each setting are indicated.

Building on this foundation, we further fine-tuned the models on task-specific datasets, employing both standard and gradual fine-tuning strategies. As illustrated in Figure 2, the gradual fine-tuning approach led to a significant boost in performance across POS tagging, NER and Keyword Identification tasks, demonstrating its effectiveness in low-resource settings.

By systematically developing and evaluating linguistically grounded resources, this work marks an important step toward enriching the NLP landscape for Mizo and Khasi languages currently transitioning from **Rising Stars** to **The Underdogs** (Joshi et al., 2020), supported by a growing suite of annotated datasets and tailored linguistic tools.

2 Related Work

2.1 POS Tagging

Cross-lingual transfer learning, as proposed by Kim et al. (2017), has been widely used for POS tagging in extremely low-resource languages by leveraging high-resource language data to improve model performance. Similarly, Chaudhary et al. (2021) introduced an active learning approach that reduces the dependency on manual annotations and mitigates conflicts in POS tag selection and optimization. More recently, Chaudhary et al. (2021) introduced the first UD-compliant POS tagging datasets for the low-resource Indic languages Angika, Magahi and Bhojpuri. Their work highlighted tokenization challenges and proposed a look-back tokenization fix that improved the F1 score, emphasizing the importance of script-aware adaptation in multilingual models. While weakly supervised POS taggers have shown promise for some low-resource languages, Kann et al. (2020) demonstrated their limitations for truly low-resource languages. The lack of good dictionaries and limited linguistic resources make traditional weak supervision methods less effective, especially for Mizo and Khasi. This highlights the need for new and better approaches.

2.2 NER & Keyword Identification

The primary challenge in NER tagging for low-resource languages is the lack of annotated data, which can be mitigated through multilingual approaches and mapping techniques. Murthy et al. (2018) demonstrated that for closely related languages, neural network layers can be divided for each language, leveraging cross-lingual features to enhance NER quality. Panchadara (2024) showed that merging datasets for Dravidian languages and utilizing mBERT and XLM-Roberta significantly improves accuracy. Dash et al. (2024) explored data augmentation techniques and community-driven resource creation to enhance NER performance for the Ho language. Similarly, Khemchandani et al. (2021) proposed RelateLM, a multilingual model that uses high-resource languages

as pivots through translation and backtranslation. Tang et al. (2019) employed an attention-based deep learning technique for clinical text classification using keyword extraction, where a fine-tuned BERT model achieved 97.6% accuracy. Bala et al. (2024) introduced a keyword extraction and summarization dataset for Mizo, enriching news articles in the language. Nasar et al. (2019) explored Keyword Identification and summarization, highlighting the lack of datasets and discussing various challenges associated with the task. These studies highlight how leveraging linguistic similarities and cross-lingual transfer can improve NER and Keyword Identification task quality for low-resource languages.

2.3 Synthetic Data Generation & Alignment

Prior studies have explored synthetic data generation using LLMs to enhance model performance Tang et al. (2023); Gholami and Omar (2023). In parallel, word alignment has been widely studied for machine translation and cross-lingual NLP Dou and Neubig (2021). Recent work by Wu et al. (2024) demonstrated the effectiveness of optimizing LLM-based models through word alignment techniques. Our work builds upon these advances by integrating synthetic data generation with word alignment techniques to improve NER performance in extremely low-resource languages.

3 Data Development

3.1 Gold Standard Data

We crawled news articles from various permitted websites in Mizo and Khasi, covering diverse topics(Healthcare, Education, Politics, Culture, Environment, Local Governance, Entertainment, and Sports) written in their respective languages. After preprocessing, we used these data to create datasets for Part-of-Speech (POS) tagging, Named Entity Recognition (NER) and Keyword Identification. These gold-standard datasets were meticulously annotated by linguistic experts with proficiency in Mizo and Khasi, ensuring high-quality and reliable annotations for downstream NLP tasks.

Due to the absence of task-specific annotation guidelines for these languages, we initially adopted the Universal Dependencies (UD) (Universal Dependencies, 2025) framework for POS tagging and later refined it to better capture their linguistic characteristics. For NER, we developed a custom annotation framework from scratch to ensure consistency and accuracy. Figure 3 shows an example of the NER dataset, and Figure 4 shows the POS dataset for both languages. We have released all the annotated datasets publicly on the iHub-Data (iHub-Data, IIIT Hyderabad, 2025) India platform[1].

Khasi: Haba ai ka jingkhublei sha ki jaitbynriew Garo ha kane ka sngi u Registrar General, High Court ka Meghalaya, u Bah E Kharumniud u la pynpaw ka jingïaroh ïa ki dkhot ka Committee jong ka Wangala kaba u la ïoh ban sakhi ïa kane ka tamasa.

1	Haba	O
2	ai	O
3	ka	O
4	jingkhublei	O
5	sha	O
6	ki	O
7	jaitbynriew	b-NEMI
8	Garo	i-NEMI
9	ha	b-NETI
10	kane	i-NETI
...		

Mizo: Chairperson thar C Zodinpuii hi Directorate of Social Welfare & Tribal Affairs-ah Joint Director-in ni 31.12.2023 ah pension in a chhuak.

1	Chairperson	b-NEMI
2	thar	O
3	C	b-NEP
4	Zodinpuii	i-NEP
5	hi	O
6	Directorate	b-NEO
7	of	i-NEO
8	Social	i-NEO
9	Welfare	i-NEO
10	&	i-NEO
...		

Figure 3: Illustration of the NER dataset with entity tags applied to the first 10 tokens of example sentences in both languages.

Inter-Annotator Agreement (IAA) Scores

Task	Khasi	Mizo
POS	0.91	0.93
NER & Keyword Identification	0.88	0.90

Table 1: Inter-Annotator Agreement (IAA) scores (Cohen's Kappa) for POS, NER and Keyword Identification datasets in Khasi and Mizo.

To validate the annotated data, we conducted an analysis of Inter-Annotator Agreement (IAA) (Artstein, 2017) using Cohen's Kappa (Rau and Shih, 2021) score. Table 1 presents Cohen's Kappa scores, and Table 2 provides dataset statistics, with a detailed breakdown for each language.

[1] https://india-data.org/datasets-listing/natural-language-processing-(nlp)/

Khasi: Haba ai ka jingkhublei sha ki jaitbynriew Garo ha kane ka sngi u Registrar General, High Court ka Meghalaya, u Bah E Kharumniud u la pynpaw ka jingïaroh ïa ki dkhot ka Committee jong ka Wangala kaba u la ïoh ban sakhi ïa kane ka tamasa.

1	Haba	CONJ
2	ai	VERB
3	ka	DET
4	jingkhublei	NOUN
5	sha	PREP
6	ki	DET
7	jaitbynriew	NOUN
8	Garo	PROPN
9	ha	PREP
10	kane	DET
...		

Mizo: Chairperson thar C Zodinpuii hi Directorate of Social Welfare & Tribal Affairs-ah Joint Director-in ni 31.12.2023 ah pension in a chhuak.

1	Chairperson	NOUN
2	thar	ADJ
3	C	PROPN
4	Zodinpuii	PROPN
5	hi	AUX
6	Directorate	NOUN
7	of	ADP
8	Social	NOUN
9	Welfare	NOUN
10	&	CCONJ
...		

Figure 4: Illustration of the POS-tagged dataset showing the first 10 tokens annotated using the adapted UD framework.

3.2 Monolingual Corpus and Synthetic Data

Using the crawled data, we compiled a monolingual corpus for each language after extensive preprocessing and filtering. The preprocessing pipeline included removal of metadata, URLs, and non-native scripts (such as Devanagari, Bengali, etc). Additionally, we applied heuristic rules for noise reduction, including filtering out texts with high proportions of negative sentiment using a sentiment classifier, and removing sentences with excessive repetition or low information density. Table 3 summarizes the final statistics of the cleaned monolingual corpora.

Additionally, we created Hindi-Mizo and Hindi-Khasi parallel datasets, using WMT23 (Pal et al., 2023) English-Mizo and English-Khasi data in conjunction with Google Translate and Bhasha-Verse (Mujadia and Sharma, 2024). To address the scarcity of annotated data further, we generated synthetic NER datasets for both languages based on the Hindi NER dataset. Figure 6 illustrates the detailed procedure for the generation of synthetic data, and Table 5 presents the statistics of these datasets.

Gold Dataset Statistics

Language	Sentences	Tokens	Types
POS Tagging			
Khasi	507	21.6K	7.5K
Mizo	502	17.3K	5.4K
NER & Keyword Identification			
Khasi	4.1K	203.1K	14.9K
Mizo	4.4K	116.2K	15.9K

Table 2: Statistics of gold-standard datasets for POS tagging, NER and Keyword Identification in Khasi and Mizo.

Monolingual Dataset Statistics

Language	Sentences	Tokens	Types
Khasi	253.3K	15.14M	269.9K
Mizo	318.4K	12.18M	294.8K

Table 3: Statistics of the Monolingual Corpora for Mizo and Khasi

4 Methodology

4.1 Baseline models

We began our experiment with baseline models, using Google MuRIL, XLM-RoBERTa-Large, and Google RemBERT. MuRIL (Khanuja et al., 2021), developed by Google, is pre-trained on 16 Indian languages. RemBERT (Chung et al., 2021), also developed by Google, is trained on 110 languages. XLM-RoBERTa-Large (Conneau et al., 2019), developed by Facebook, is pre-trained on 100 languages.

For all three tasks and both languages, we first applied a zero-shot approach to the gold-standard data. For Mizo, XLM-RoBERTa-Large achieved the best performance in both POS tagging and NER. For Khasi, RemBERT performed best for POS tagging, while XLM-RoBERTa-Large was the top performer for NER and Keyword Identification. Table 4 presents the detailed results of our baseline models.

4.2 Model Finetune

As these models perform poorly in a zero-shot setting, a two-stage fine-tuning approach is adopted. In the first stage, the models are fine-tuned on a monolingual corpus to enhance their understanding of the target languages. Once language compre-

F1 Scores of Baseline Models

Language	MuRIL	RemBERT	XLM-R large
POS Tagging			
Khasi	9.47	14.19	11.62
Mizo	12.94	9.11	17.38
NER & Keyword Identification			
Khasi	8.59	9.31	16.28
Mizo	12.35	8.61	13.07

Table 4: Macro F1-scores for POS tagging, NER, and Keyword Identification in a zero-shot setting using baseline models.

hension is established, task-specific fine-tuning is performed. Two setups are explored: Standard and Gradual. Section 6 provides a detailed explanation of this process, while Table 7 presents the corresponding results.

5 Synthetic NER Data Generation

There is a severe lack of publicly available data for these languages on the internet, making it necessary to rely on synthetic data generation (Anonymous, 2025) to obtain large-scale resources without direct human involvement. However, direct translation from another language is not feasible, as it often results in variations in word count and word order (James and Krishnamurthy, 2025). This makes it difficult to map the NER tags, especially when using the BIO (Beginning, Inside, Outside) (Yohannes and Amagasa, 2022) format.

To address this, we used Hindi NER (Bahad et al., 2024) data (tagged in BIO format) as our source. We first translated the sentences without their tags into Mizo and Khasi (P M et al., 2024). After translation, we aligned the words using Awesome-Align and VecMAP.

- **Awesome-Align** (Dou and Neubig, 2021) is a cross-lingual word alignment tool that leverages multilingual BERT (mBERT) to generate high-quality word alignments between parallel texts.
- **VecMAP** (Artetxe et al., 2018, 2017) is a method for learning cross-lingual word embeddings by mapping word vectors from one language to another into a shared vector space, allowing better alignment and improving translation consistency.

Hindi: वे यहोशू के पास लौट आए।
(ve yahoshoo ke paas laut aae.)
Mizo: Josua hnênah an kîr leh a .

0-2 || वे → an
1-0 || यहोशू → Josua
2-1 || के → hnênah
3-1 || पास → hnênah
4-3 || लौट → kîr
5-5 || आए → a

Hindi: उसने अपनी आँखों से मुझे धन्यवाद दिया।
(usane apanee aankhon se mujhe dhanyavaad diya.)
Khasi: U khublei ianga da ki khmat.

0-0 || उसने → U
1-6 || अपनी → jongu
2-5 || आँखों → khmat
3-3 || से → da
4-2 || मुझे → ianga
5-1 || धन्यवाद → khublei
6-1 || दिया → khublei

Figure 5: Detailed alignment examples for Hindi–Khasi and Hindi–Mizo translations after refinement using Awesome-Align and VecMAP. Each example includes the original Hindi sentence, its transliteration, the corresponding target translation (Mizo & Khasi), and word-level alignments.

To train Awesome-Align, we utilized the WMT23 English-Mizo and English-Khasi parallel datasets (Pal et al., 2023). Since our source data was in Hindi, we first translated the English sentences into Hindi. Subsequently, we trained Awesome-Align using the Hindi-Mizo and Hindi-Khasi parallel datasets.

However, Awesome-Align internally relies on mBERT (Devlin et al., 2018), which has minimal to no representation of Mizo and Khasi. To mitigate this limitation, we first fine-tuned mBERT on our monolingual corpus. The initial results were suboptimal, prompting us to refine our approach. We partitioned the monolingual corpus into two subsets, each containing approximately 7.5 million tokens. The model was initially fine-tuned on the first subset, followed by an additional fine-tuning stage on the second subset. This two-stage fine-tuning process resulted in a perplexity score of 9.25, significantly enhancing the model's ability to process Mizo and Khasi text.

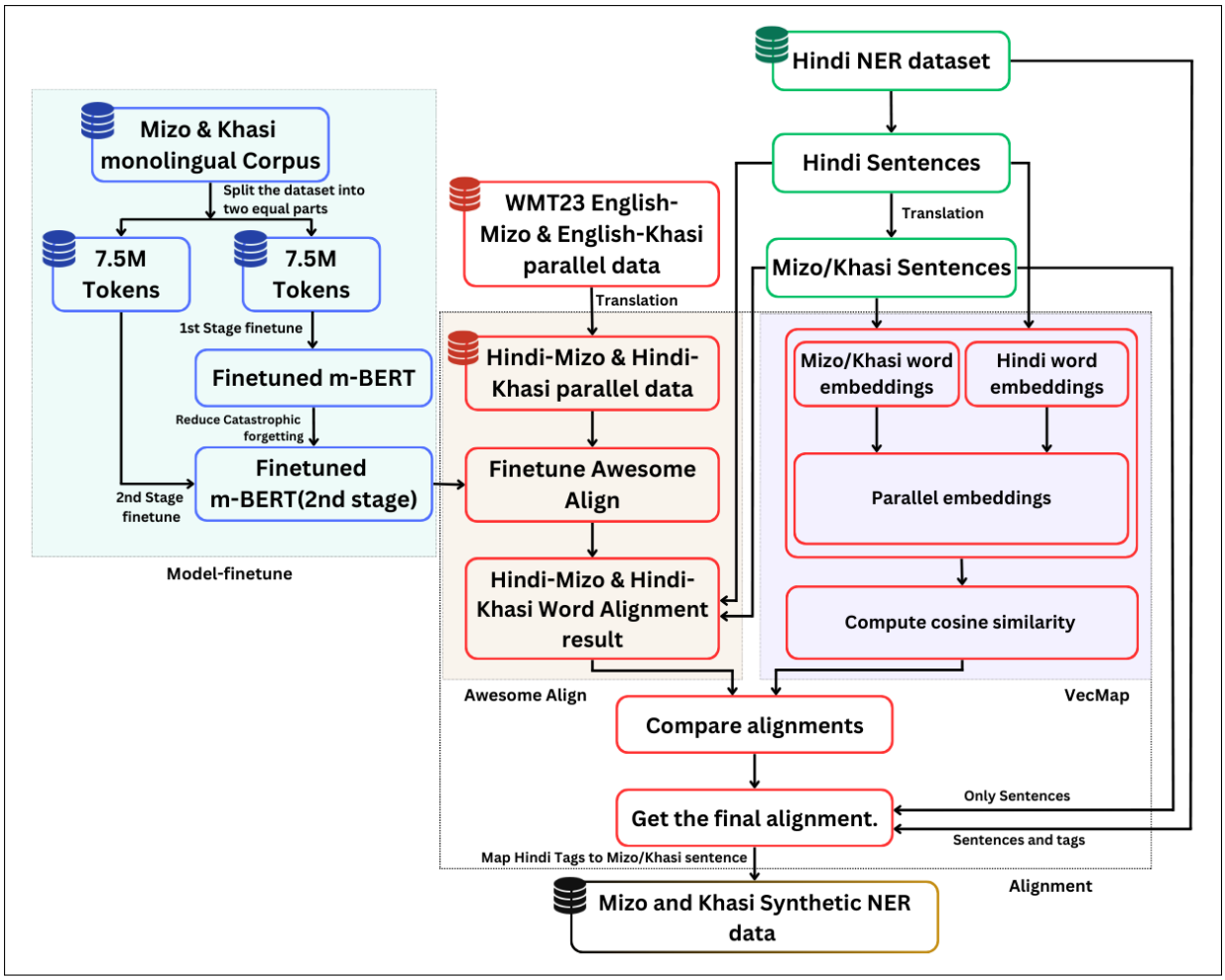

Figure 6: Pipeline for synthetic NER data generation.

Once Awesome-Align was trained, we used our Hindi source sentences and their Mizo/Khasi translations (Hindi ||| Mizo/Khasi) to generate word alignments. However, the model occasionally produced unaligned words or incorrectly mapped multiple words to a single word. To refine these alignments, we used VecMAP.

For VecMAP, we first generated Word2Vec (Mikolov et al., 2013) embeddings for Hindi, Mizo, and Khasi using our source Hindi sentences and their corresponding translations. We then mapped the Hindi embeddings to a common space with Mizo/Khasi embeddings and vice versa. Using cosine similarity, we corrected the unaligned words and improved alignments where Awesome-Align incorrectly assigned multiple words to a single word. This resulted in more accurate alignments. Figure 5 illustrates detailed examples of Hindi–Khasi and Hindi–Mizo alignments.

At this stage, we had Hindi NER data, along with translated Mizo/Khasi sentences and their word alignments. To map the NER tags, we first removed the BIO tags and then assigned the tags according to the alignments. Finally, we reapplied the BIO tags:

- **B (Beginning)** was assigned to the first token of an entity.
- **I (Inside)** was assigned to subsequent tokens within the entity.
- **O (Outside)** was assigned to tokens that did not belong to any entity.

This process allowed us to generate high-quality synthetic NER data for Mizo and Khasi, ensuring accurate tag mappings despite the complexities of translation and word alignment. Figure 6 illustrates the detailed procedure for synthetic data generation, and Table 5 presents the statistics of these datasets. All synthetic datasets have been publicly released on the iHub-Data India platform[2].

[2] https://india-data.org/datasets-listing

Synthetic NER data

Dataset	Sentences	Tokens	Types
Khasi	6.6K	220.3K	15.1K
Mizo	6.6K	175.2K	17.4K

Table 5: Statistics of the synthetic NER dataset for Mizo and Khasi.

6 Experiments and Results

6.1 1st Stage Finetune

While these models support several Indian languages and scripts, they do not accommodate Mizo and Khasi, as no datasets for these languages were included during pre-training. Although their vocabularies contain the Latin script, which is also used by Mizo and Khasi, the structural differences in these languages limit the models' ability to understand them effectively. Consequently, their zero-shot performance on Mizo and Khasi was significantly low.

Perplexity Scores from First-Stage Fine-Tuning

Language	MuRIL	RemBERT	XLM-R large
Khasi	5.19	8.13	8.57
Mizo	10.06	7.69	7.92

Table 6: Perplexity scores after the first stage of fine-tuning on the monolingual corpus.

To address this limitation, we fine-tuned these models on a monolingual corpus specifically curated for Mizo and Khasi. This fine-tuning process improved their language comprehension, making them more suitable for downstream NLP tasks. We evaluated the effectiveness of this adaptation using perplexity scores, with detailed results presented in Table 6.

6.2 2nd Stage Finetune (Task-Specific)

With these models now adapted to our target languages, they are ready for fine-tuning on specific NLP tasks. For each task, we employ two fine-tuning strategies: standard fine-tuning and gradual fine-tuning. In gradual training, we initially freeze all model layers and progressively unfreeze them over several epochs. Using these approaches, we achieved an F1 score improvement of approximately 62% for POS and 43% for NER and Keyword Identification in the standard fine-tuning setup, with an additional gain of 6% when applying gradual training.

6.2.1 POS Tagging

Part-of-Speech POS tagging involves labeling each word in a sentence with its corresponding grammatical categories such as noun, verb, adjective, or adverb. Building on our first-stage fine-tuned model, we further fine-tuned it on our gold-standard POS tagging dataset and evaluated its performance on the same dataset. In the standard fine-tuning setup, MuRIL performed slightly better for Mizo, while RemBERT yielded the best results for Khasi. However, with gradual training, MuRIL achieved the highest performance for Khasi, whereas XLM-RoBERTa-Large outperformed other models for Mizo. The detailed results are presented in Table 7.

6.2.2 NER Tagging & Keyword Identification

Named Entity Recognition (NER) involves extracting meaningful information from text by identifying and categorizing named entities such as person names, locations, and organizations. Additionally, tasks beyond NER, Keyword Identification, focus on extracting key terms that represent the main topics of a document. This is particularly useful for applications like search engine optimization, text summarization, and content classification.

To evaluate NER performance, we fine-tuned our first-stage fine-tuned models on synthetically generated NER data and used gold-standard data as a benchmark. In the standard fine-tuning setup, XLM-RoBERTa-Large achieved the best performance for Khasi, while MuRIL performed better for Mizo. However, in the gradual fine-tuning setup, MuRIL outperformed other models for Khasi, while it remained the best-performing model for Mizo. The detailed results are presented in Table 7.

7 Conclusion

The development of NLP resources for low-resource languages such as Mizo and Khasi is crucial for their digital preservation and broader linguistic accessibility. Through the creation of high-quality annotated datasets for POS tagging, NER, and Keyword Identification, this work establishes foundational linguistic resources to support future research and tool development for these underrepresented languages. In particular, our synthetic NER data generation pipeline leveraging translation and word alignment demonstrate the feasibility

F1 Score of Task-Specific Fine-Tuning Across Different Models

Language	Standard			Gradual		
	MuRIL	RemBERT	XLM-R-Large	MuRIL	RemBERT	XLM-R-Large
POS tagging						
Khasi	76.49	82.51	71.02	**83.52**	81.15	76.81
Mizo	79.53	73.26	75.41	81.35	79.60	**82.39**
NER and Keyword Identification						
Khasi	48.30	47.11	51.68	**57.84**	55.27	53.79
Mizo	61.88	58.69	59.27	**66.79**	64.08	64.92

Table 7: Macro F1 score comparison of fine-tuned MuRIL, RemBERT, and XLM-R Large on POS tagging and NER and Keyword Identification tasks for Mizo and Khasi under standard and gradual fine-tuning setups.

of bootstrapping annotated data in the absence of gold-standard resources.

Among the models evaluated, MuRIL and XLM-R Large emerged as the most effective choices, depending on the task. MuRIL performed best for Khasi POS tagging (f1: 83.52) and both Mizo NER (f1:66.79) and Khasi NER (f1:57.84), while XLM-R Large achieved the highest score (f1:82.39) for Mizo POS tagging, demonstrating how a well-structured fine-tuning strategy can significantly enhance model performance.

Future work can extend these efforts by expanding annotated datasets, refining task-specific guidelines, and increasing coverage across linguistic phenomena. Incorporating community-driven or semi-automated annotation strategies may further enhance the scalability and adaptability of resource creation, contributing to better representation and accessibility for Mizo, Khasi, and other low-resource languages.

Acknowledgment

This project is funded by I-Hub Data (iHub-Data, IIIT Hyderabad, 2025). We thank I-Hub Data as well for providing access to computational resources (GPU support), which enabled us to carry out the training and experiments effectively. We are grateful to Nagaraju Vuppala and Mayuri Dilip for their contributions to data management in the initial stages. Finally, we extend our sincere thanks to all anonymous reviewers for their valuable feedback and constructive suggestions, which helped improve the quality of this work.

Ethics Statement

This research promotes linguistic inclusivity by developing NLP resources for Mizo and Khasi, two extremely low-resource languages. Dataset creation involved collaboration with native speakers and language experts, ensuring ethical data collection and annotation while respecting linguistic and cultural contexts.

Textual data was collected from permitted news websites in full compliance with their terms of use. All human annotators participated voluntarily and were fairly remunerated for their work. The dataset contains no personally identifiable information, ensuring privacy and confidentiality.

While multilingual models were fine-tuned on these languages, potential biases remain due to the limited availability of digital resources. We encourage further community-driven efforts to enhance NLP for underrepresented languages.

We used LLM to refine sentence structure and check grammar in our paper, ensuring clarity while maintaining the originality of the content.

References

Anonymous. 2025. Does synthetic data help named entity recognition for low-resource languages? In *Submitted to ACL Rolling Review - December 2024*. Under review.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

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*.

Ron Artstein. 2017. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.

Sankalp Bahad, Pruthwik Mishra, Parameswari Krishnamurthy, and Dipti Sharma. 2024. Fine-tuning pre-trained named entity recognition models for Indian

languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 75–82, Mexico City, Mexico. Association for Computational Linguistics.

Abhinaba Bala, Ashok Urlana, Rahul Mishra, and Parameswari Krishnamurthy. 2024. Exploring news summarization and enrichment in a highly resource-scarce indian language: A case study of mizo. *arXiv preprint arXiv:2405.00717*.

Census Commissioner. 2022. Census of India 2011 - Language Atlas. [Accessed 05-06-2024].

Aditi Chaudhary, Antonios Anastasopoulos, Zaid Sheikh, and Graham Neubig. 2021. Reducing confusion in active learning for part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 9.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Satya Ranjan Dash, Bikram Biruli, Yasobanta Das, Prosper Abel Mgimwa, Muhammed Abdur Rahmaan Kamaldeen, and Aloka Fernando. 2024. Named entity recognition (ner) in low resource languages of ho. In *Empowering Low-Resource Languages With NLP Solutions*, pages 157–182. IGI Global.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*.

Sia Gholami and Marwan Omar. 2023. Does synthetic data make large language models more efficient? *arXiv preprint arXiv:2310.07830*.

iHub-Data, IIIT Hyderabad. 2025. iHub-Data IIIT Hyderabad. Accessed: 9 Mar. 2025.

Indian-Constitution. 2022. Languages included in the eighth schedule of the indian constitution. [Accessed: 2 Mar. 2025].

Antony Alexander James and Parameswari Krishnamurthy. 2025. Pos-aware neural approaches for word alignment in dravidian languages. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 154–159.

M. Jenny and P. Sidwell. 2014. *The Handbook of Austroasiatic Languages (2 vols)*. Grammars and Sketches of the World's Languages. Brill.

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*.

Katharina Kann, Ophélie Lacroix, and Anders Søgaard. 2020. Weakly supervised pos taggers perform poorly on truly low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8066–8073.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: Multilingual representations for indian languages. *arXiv preprint arXiv:2103.10730*.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting language relatedness for low web-resource language model adaptation: An indic languages study. *arXiv preprint arXiv:2106.03958*.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for pos tagging without cross-lingual resources. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2832–2838.

Sanjeev Kumar, Preethi Jyothi, and Pushpak Bhattacharyya. 2024. Part-of-speech tagging for extremely low-resource Indian languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14422–14431, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Vandan Mujadia and Dipti Misra Sharma. 2024. Bhashaverse: Translation ecosystem for indian subcontinent languages. *arXiv preprint arXiv:2412.04351*.

Rudra Murthy, Mitesh M Khapra, and Pushpak Bhattacharyya. 2018. Improving ner tagging performance in low-resource languages via multilingual learning. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 18(2):1–20.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik. 2019. Textual keyword extraction and summarization: State-of-the-art. *Information Processing & Management*, 56(6):102088.

Abhinav P M, Ketaki Shetye, and Parameswari Krishnamurthy. 2024. MTNLP-IIITH: Machine translation for low-resource Indic languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 751–755, Miami, Florida, USA. Association for Computational Linguistics.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. Findings of the WMT 2023 shared task on low-resource Indic language translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.

Kiranmaye Panchadara. 2024. Enhancing named entity recognition in low-resource dravidian languages: A comparative analysis of multilingual learning and transfer learning techniques. *Journal of Artificial intelligence and Machine Learning*, 2(1):1–7.

Gerald Rau and Yu-Shan Shih. 2021. Evaluation of cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. *Journal of english for academic purposes*, 53:101026.

Sunita Sarkar, Sneha Das, Basab Nath, and Somnath Mukhopadhyay. 2024. A multilingual neural machine translation model for low resource north eastern languages.

Matthew Tang, Priyanka Gandhi, Md Ahsanul Kabir, Christopher Zou, Jordyn Blakey, and Xiao Luo. 2019. Progress notes classification and keyword extraction using attention-based deep learning models with bert. *arXiv preprint arXiv:1910.05786*.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *arXiv preprint arXiv:2303.04360*.

G. Thurgood and R.J. LaPolla. 2003. *The Sino-Tibetan Languages*. Routledge language family series. Routledge.

Universal Dependencies. 2025. Universal POS tags. Accessed: 7 March 2025.

Qiyu Wu, Masaaki Nagata, Zhongtao Miao, and Yoshimasa Tsuruoka. 2024. Word alignment as preference for machine translation. *arXiv preprint arXiv:2405.09223*.

Hailemariam Mehari Yohannes and Toshiyuki Amagasa. 2022. Named-entity recognition for a low-resource language using pre-trained language model. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, page 837–844, New York, NY, USA. Association for Computing Machinery.

A Appendix

A.1 Linguistic Landscape of Mizo

Mizo, a Tibeto-Burman language (Thurgood and LaPolla, 2003), is written in the Roman script, which was introduced by Welsh Christian missionaries in the late 19th century. The early Mizo script was developed by Rev. J.H. Lorrain and Rev. F.W. Savidge in 1894. The Mizo alphabet consists of 25 letters, excluding F, Q, R, and X, as these letters do not exist in native Mizo words.

Beyond being a means of communication, Mizo serves as a symbol of identity, unity, and cultural heritage for the Zo people.It is **spoken by** approximately **831K** (according to Census 2011) people in India and is primarily used in Mizoram (Fig. 7). Additionally, Mizo (or closely related dialects) is spoken in parts of Manipur, Tripura, Assam, as well as in neighboring Myanmar and Bangladesh, where different Zo communities reside.

Figure 7: Map of India highlighting Mizoram[3], the primary region where Mizo is spoken.

Mizo evolved from various dialects spoken by different Zo tribes. Historically, the Lusei dialect (spoken by the Lusei/Lushai tribe) became dominant due to its early adoption in education, administration, and Christian missionary work. Over time, other dialects merged into what is now recognized as the standard Mizo language. However, distinct Zo dialects such as Hmar, Paite, Lai, Mara, and Vaiphei continue to be spoken by their respective communities.

[3]Source: https://tinyurl.com/5b6893an

Linguistically, Mizo is an agglutinative language, meaning words are formed by adding multiple affixes to a root word, allowing complex meanings to be expressed through morphological constructions rather than separate words..

A.2 Linguistic Landscape Khasi

Khasi belongs to the Austroasiatic language family (Jenny and Sidwell, 2014) and is predominantly spoken in Meghalaya, India, with approximately **1.4 million speakers** (according to Census 2011). It is written in the Roman script and has a rich oral tradition.

Khasi is the largest indigenous language in Meghalaya (Fig: 8) and is primarily spoken in the Khasi and Jaintia Hills, as well as the Ri Bhoi district. The Khasi people are linked to the Mon-Khmer sub-group of the Austroasiatic language family, with linguistic similarities to Mon-Khmer dialects spoken in Southeast Asia.

Figure 8: Map of India highlighting Meghalaya[4], the primary region where Khasi is spoken.

Historically, the Khasi people are known as Hynniewtrep (Children of Seven Huts), representing seven sub-groups: Khynriam, Pnar (Jaintia), Bhoi, War, Maram, Lyngngam, and Mnar. Among these, the Pnar (Jaintia), Bhoi, and War are significant regional variations. While Khasi has a standardized written form, dialectal variations exist across different regions.

[4]Source: https://tinyurl.com/5fyebpp3

Linguistically, Khasi is an agglutinative language, where words are formed by adding prefixes, suffixes, and infixes to a root word, allowing complex meanings to be built through morphological processes rather than separate words. .

A.3 Experimental Setup

Multilingual transformer-based models, including MuRIL, RemBERT, and XLM-RoBERTa-Large, were fine-tuned on Mizo and Khasi datasets. The models were initialized with pre-trained weights and further trained using our annotated datasets. Fine-tuning was conducted using the Hugging Face Transformers library on NVIDIA L40S GPU (96GB VRAM). The training process followed a two-stage fine-tuning approach:

- **Stage 1 (Monolingual Fine-Tuning)**
  - Batch size: **32**
  - Learning rate: **3e-5**
  - Epochs: **2**
- **Stage 2 (Task-Specific Fine-Tuning for NER/POS)**
  - Batch size: **16**
  - Learning rate: **2e-5**
  - Epochs: **3**

For optimization, the **AdamW** optimizer was used with a **linear decay learning rate schedule**.