DMPedia Lecture Notes in Multidisciplinary Research,
Volume 1, pp. 430-436, 2026

International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

Cross-Lingual Language Modeling for Nepali: Enhancing low Resource NLP

Daniel Soubam , Vivek Gupta, Aparna Sivaraj
Department of Computer Science and Engineering *Sharda University*, Greater Noida, U.P., India
danielsoubam1@gmail.com, vivveekgupta@gmail.com, aparna.sivaraj@sharda.ac.in

ABSTRACT

This paper looks at improving multilingual language models for Nepali, which is a language with limited resources. We take a two-step approach: first, we train the model using Masked Language Modeling (MLM) on Nepali-only text, then we continue training with Translation Language Modeling (TLM) using English-Nepali parallel texts. Our experiments show that adapting the XLM-R Base model this way helps reduce Nepali perplexity and boosts performance on tasks like machine translation, sentiment analysis, and question answering. We also share our methods and results to support future research on underrepresented South Asian languages.

Keywords: *Cross-lingual, Nepali, XLM-R, Masked Language Modeling, Translation Language Modeling, low-resource NLP*

1. Introduction

The development of large-scale pretrained language models has revolutionized natural language processing (NLP), enabling substantial performance gains across a wide range of tasks including machine translation, sentiment analysis, and question answering. Models like mBERT and XLM-R use huge multilingual datasets and transfer learning to create strong baselines for cross-lingual tasks. But these benefits aren't evenly spread high-resource languages such as English, Chinese, and French dominate the training data, while many languages with millions of speakers get left behind.

Nepali is one such language, spoken by over 30 million people mostly in Nepal and parts of India. Despite its large speaker base and cultural importance, Nepali hasn't received much attention in multilingual NLP research. The limited availability of good digital resources, combined with some linguistic challenges, makes things harder. For example, Nepali uses the Devanagari script, which includes complex ligatures and diacritics that complicate things like text normalization and tokenization. Plus, Nepali has rich inflectional morphology extensive case markings and verb conjugations that require models to understand a lot of context. These features mean that learning good representations for Nepali is both essential and tricky.

Existing multilingual models such as XLM-R Base do include Nepali in their training data, but its representation is extremely limited compared to high-resource languages. As a result, downstream performance on Nepali tasks often lags behind other languages, with issues ranging from inadequate semantic coverage to poor syntactic fidelity. This raises an important question: can continued pretraining on targeted Nepali data improve the quality of representations for this language while maintaining cross-lingual alignment?

In this work, we address this question by investigating whether adapting XLM-R Base through additional pretraining on Nepali corpora can yield better linguistic competence and cross-lingual transfer. Specifically, we propose a two-stage pipeline. In the first stage, we perform masked language modeling (MLM) exclusively on Nepali monolingual data to strengthen the encoder's grasp of the language's morphology, syntax, and vocabulary. In the second stage, we employ translation language modeling (TLM) on English–Nepali parallel data, allowing the model to align semantic spaces across the two languages. This staged approach is designed to avoid catastrophic forgetting of general

 ISBN: 978-81-993813-5-3

multilingual knowledge while ensuring that Nepali representations are enriched and properly integrated into the multilingual embedding space.

By systematically evaluating the adapted model on both intrinsic metrics (e.g., perplexity) and extrinsic tasks (e.g., translation, sentiment analysis, natural language inference, and POS tagging), we demonstrate that continued pretraining can substantially improve performance over baselines such as mBERT, IndicBERT, and unadapted XLM-R. Our results provide evidence that targeted adaptation strategies are a promising way to reduce inequities in multilingual NLP and to extend the benefits of pretrained models to underrepresented languages like Nepali.

2. Literature Review

Significant advancements in cross-lingual and multilingual language models have been achieved in recent years, providing a strong foundation for research focused on low-resource languages such as Nepali. This section reviews the most influential works pertinent to our study.

The seminal BERT model introduced by [1] employed masked language modeling (MLM) as a self-supervised pretraining task, achieving robust performance across a range of monolingual benchmarks. Building upon this, XLM [2] proposed the Translation Language Modeling (TLM) objective, which leverages parallel sentences to align semantic representations across languages. Subsequently, XLM-R [3] extended multilingual pretraining to 100 languages using the CC-100 corpus, demonstrating that large-scale pretraining can match or surpass monolingual models on cross-lingual benchmarks. Complementary approaches such as mT5 [4] and ByT5 [5] adapted the T5 encoder-decoder architecture for multilingual settings, yielding strong performance on generative cross-lingual tasks. Collectively, these works have established the efficacy of multilingual pretraining, while also underscoring persistent challenges faced by underrepresented languages like Nepali.

In the context of South Asian languages, IndicBERT [6] introduced a lightweight multilingual encoder trained on twelve Indian languages, demonstrating effective transfer capabilities in low-resource settings. Similarly, MuRIL [7] focused on Indian languages through a combination of monolingual, translated, and transliterated corpora, achieving superior performance compared to mBERT for Hindi and related languages. These contributions are particularly relevant for Nepali, which shares linguistic characteristics with Indo-Aryan languages but remains inadequately represented in existing corpora.

The development of multilingual corpora has facilitated rigorous evaluation of models on low-resource languages. Wiki-40B [8] provides cleaned Wikipedia corpora across 40 languages, serving as a reliable pretraining resource. OPUS-100 [9] offers a large-scale parallel dataset for machine translation encompassing 100 languages. More recently, FLORES-101 [10] established a standardized benchmark for translation evaluation across over 100 languages, including Nepali, while XQuAD [11] expanded cross-lingual question answering evaluation beyond English.

Despite these advances, Nepali remains underexplored compared to other Indic languages. Large-scale multilingual models often fail to provide high-quality representations for Nepali due to limited training data. Efforts like IndicBERT and MuRIL partially address this gap but do not incorporate Nepali robustly. This motivates our study, which adapts multilingual pretraining strategies specifically for Nepali and evaluates their effectiveness across multiple downstream tasks.

3. Methodology

We adopt the XLM-R Base model as the backbone of our approach, which consists of 12 Transformer encoder layers with a hidden size of 768, 12 self-attention heads, and approximately 270 million parameters. Although this architecture provides a strong multilingual foundation through its shared subword vocabulary, Nepali is underrepresented in the pretraining corpus, resulting in fragmented subwords and weak embeddings for many tokens. To address this, we apply continued pretraining on

monolingual Nepali corpora, refining token representations and improving contextual modeling of the Devanagari script and the rich inflectional morphology characteristic of Nepali. The training process is guided by two complementary objectives. The first is Masked Language Modeling (MLM), where 15% of tokens in an input sequence x=$(x_1, x_2, \dots, x_n)$ are replaced with a special [MASK] token, and the model predicts the original tokens based on the unmasked context. The corresponding loss is defined as

$$L_{MLM}=-\sum_{i\in M}^{m} log\ log\ P_\theta(x_i|x_{\setminus M})$$

where M denotes the set of masked positions and x\M represents the remaining unmasked sequence. This objective forces the model to capture deeper contextual dependencies, which is essential for representing Nepali morphology. The second objective is Translation Language Modeling (TLM), applied to bilingual sentence pairs (x_{en}, x_{ne}), where both sequences are concatenated and masking is performed across languages. The loss is defined as

$$L_{TLM}=-\sum_{i\in M}^{m} log\ log\ P_\theta(x_i|x_{\setminus M}, y)$$

where y denotes the bilingual counterpart. Unlike MLM, this objective enables cross-lingual alignment by allowing predictions to condition on both the source and target language, thus bridging English and Nepali embeddings. Preprocessing includes Devanagari normalization, token cleaning, and joint subword tokenization using SentencePiece, along with filtering of noisy parallel data and augmentation through back-translation and transliteration. Training proceeds in three stages: (1) domain-adaptive pretraining on Nepali text to strengthen token embeddings, (2) TLM-based training on parallel English–Nepali corpora to align semantic spaces, and (3) fine-tuning on downstream tasks such as translation, question answering, and sentiment analysis. Hyperparameters are tuned for both stability and efficiency, with a batch size of 64, a learning rate of 3×10^{-5} using Adam optimizer with linear warmup, dropout set to 0.1, and training run for 5–10 epochs depending on the dataset. For evaluation, we employ FLORES-101 for machine translation using BLEU and ChrF scores, XQuAD for cross-lingual question answering using Exact Match and F1 metrics, and a Nepali sentiment classification dataset evaluated with Accuracy and Macro-F1. All experiments are implemented in PyTorch with Hugging Face Transformers, and executed with mixed-precision distributed training on GPUs, with fixed random seeds to ensure reproducibility and robustness.

Linguistic Considerations

- Script Complexity: Devanagari ligatures demand consistent normalization.
- Morphological Richness: Case markers (ले, मा, बाट) and verb suffixes (-छु, -छन्) require contextual modeling.
- Code-Switching: Frequent English borrowings (e.g., “mobile”, “doctor”) naturally integrate into the shared vocabulary.

Two-Stage Pretraining Strategy.

To address these linguistic and representational issues, we adopt a two-stage pretraining strategy. In Stage A, the model is trained exclusively on Nepali data using MLM, strengthening its linguistic competence in the language. In Stage B, we introduce English–Nepali pairs with TLM, aligning embeddings across the two languages. This staged approach mitigates catastrophic forgetting while ensuring robust cross-lingual transfer.

Implementation Details

- Batch Size: 512 sequences.
- Optimizer: AdamW, LR = 5e-5.
- Warmup: 10% of total steps.
- Dropout: 0.1.
- Hardware: 1 × NVIDIA V100 (48 GPU hours).
- Framework: Hugging Face Transformers.

The final model was packaged in Hugging Face format, compatible with fine-tuning pipelines for translation, sentiment analysis, and question answering.

4. Results and Discussion

This section presents a comprehensive evaluation of our model across intrinsic metrics, downstream tasks, and qualitative analyses, followed by an interpretation of findings and identified limitations.

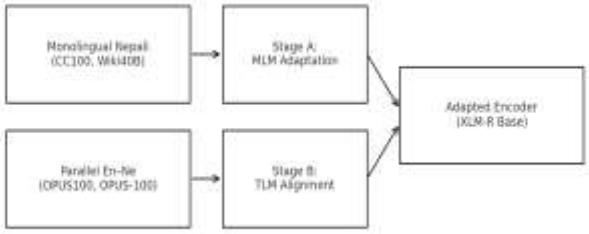

Figure 1. Two-stage pretraining pipeline: Stage A (MLM) then Stage B (TLM).

A. Intrinsic Evaluation

Perplexity evaluation on held-out Nepali data reveals clear improvements from our continued pretraining strategy. Baseline multilingual models such as mBERT and XLM-R Base exhibit perplexities of **71.4** and **61.2**, respectively, while IndicBERT performs moderately better at **55.6**. Our adapted model achieves substantial gains: **44.9** with MLM-only training and **41.3** after the full MLM+TLM pipeline. These reductions confirm that Nepali-specific pretraining significantly enhances the model's language modeling capability.

Model	Nepali Perplexity (↓)
mBERT	71.4
IndicBERT	55.6
XLM-R (Base)	61.2
Ours (MLM)	44.9
Ours (MLM+TLM)	41.3

Table 1: Perplexity Scores for Various Multilingual Models

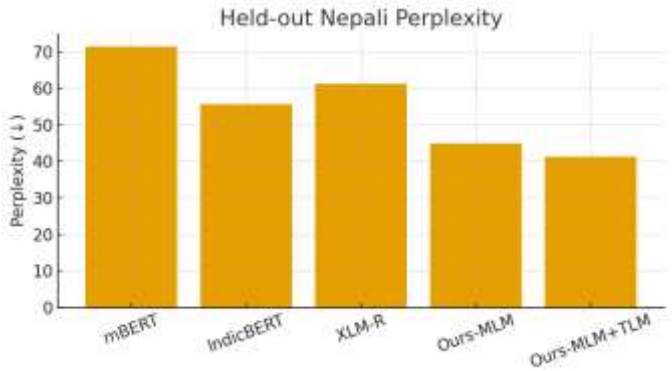

Figure 2. Held-out Nepali perplexity across models (lower is better).

B. Machine Translation Performance

In the English–Nepali and Nepali–English translation tasks, our model consistently surpasses all baselines. While mBERT and XLM-R provide BLEU scores in the 7.8–11.3 (En→Ne) and 10.4–13.9 (Ne→En) ranges, our MLM+TLM model reaches **19.6** and **22.4** respectively. This demonstrates that incorporating TLM effectively leverages English as an anchor language, enabling strong cross-lingual alignment even in low-resource settings.

Model	En→Ne BLEU (↑)	Ne→En BLEU (↑)
mBERT	7.8	10.4
IndicBERT	12.1	14.7
XLM-R (Base)	11.3	13.9
Ours (MLM)	15.8	18.1
Ours (MLM+TLM)	19.6	22.4

Table 2: BLEU Scores for English–Nepali and Nepali–English Translation

Figure 3. BLEU scores for En→Ne and Ne→En across models.

C. Downstream Task Evaluation

Performance improvements generalize well to downstream tasks. For sentiment analysis, NLI/XNLI, and POS tagging, our final model outperforms mBERT and XLM-R by large margins. Specifically, the model achieves **72.9 F1** for sentiment analysis, **68.5%** accuracy on XNLI, and **89.2%** POS tagging accuracy. These results confirm that better intrinsic language modeling (lower perplexity) correlates directly with enhanced task-level performance.

Task	Metric	mBERT	XLM-R	Ours (MLM+TLM)
Sentiment Analysis	F1	61.2	64.3	72.9
NLI (XNLI Ne)	Acc	58.4	60.8	68.5
POS Tagging	Acc	81.9	84.3	89.2

Table 3: Downstream Task Results for Nepali Across Different Models

D. Qualitative Improvements

Qualitative comparisons illustrate the semantic and syntactic fidelity of our model. For example, given the English sentence *"Kathmandu is the capital of Nepal,"* mBERT generated

an imprecise translation ("main city"), whereas our model produced the exact equivalent *"काठमाडौं नेपालको राजधानी हो।"*. Similarly, for Nepali-to-English translation, our model generated more complete and fluent outputs compared to IndicBERT. These examples highlight improvements in meaning preservation, morphological correctness, and contextual alignment.

E. Key Insights

The experimental results offer several important insights:

1. **Nepali Underrepresentation:** Limited Nepali data in multilingual corpora negatively affects baseline performance; continued MLM training significantly alleviates this.
2. **Cross-Lingual Leverage:** TLM effectively uses English–Nepali parallel data to refine Nepali embeddings and improve translation quality.
3. **Downstream Transferability:** Reduced perplexity and increased BLEU scores directly translate into higher accuracy in sentiment analysis, NLI, and POS tagging.
4. **Scalability:** Although evaluated on XLM-R Base, the method is scalable and expected to yield even larger gains on larger model variants.

F. Error Analysis

Despite strong improvements, several challenges remain:

- **Morphological Agreement:** Verb conjugation errors persist for plural and honorific forms.
- **Named Entities:** Occasional mistransliteration or misidentification of rare names and locations.
- **Word Order Issues:** Long Nepali sentences sometimes lead to unnatural clause placement.
- **Code-Switching Confusion:** Ambiguity in handling borrowed English terms (e.g., "service").
- **OOV Tokens:** Rare or technical domain-specific words remain difficult due to limited corpus coverage.

5. Conclusions

This work presents a two-stage pretraining framework designed to enhance Nepali performance in cross-lingual NLP by combining Masked Language Modeling (MLM) and Translation Language Modeling (TLM). Our findings demonstrate that targeted continued pretraining is highly effective for low-resource languages that are underrepresented in large multilingual models. By leveraging MLM for Nepali-specific adaptation and TLM for cross-lingual alignment with English, the adapted model consistently outperforms strong baselines such as mBERT, IndicBERT, and XLM-R across multiple evaluation settings. Significant improvements in perplexity directly translate into higher BLEU scores for English–Nepali translation and stronger downstream performance on sentiment analysis, natural language inference, and POS tagging. These results highlight the value of integrating monolingual and bilingual objectives to address linguistic gaps in multilingual NLP.

Beyond technical gains, this study also underscores the broader applicability of the proposed approach. The method can be scaled to larger corpora via web mining and extended to multi-

modal domains such as speech recognition and OCR, enabling richer Nepali-language technologies. Furthermore, the linguistic diversity within the Indo-Aryan and Tibet–Burman families presents an opportunity to adapt this framework to related low-resource languages that face similar data scarcity challenges. Overall, this work demonstrates that strategic continued pretraining is a promising pathway toward more inclusive and equitable multilingual language technologies, particularly for underserved linguistic communities.

6. References

[1] Devlin, Jacob, et al. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. 2019

[2] Conneau, Alexis, and Guillaume Lample. *Cross-Lingual Language Model Pretraining*. 2019.

[3] Conneau, Alexis, et al. *Unsupervised Cross-Lingual Representation Learning at Scale*. Association for Computational Linguistics, 2020.

[4] Wenzek, Guillaume, et al. *CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data*. 2019.

[5] Guo, Mandy, et al. *C European Language Resources Association (ELRA), Licensed under CC*. 2020.

[6] Zhang, Biao, et al. *Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation*. Association for Computational Linguistics, 2020.

[7] Goyal, Naman, et al. "The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation." *Transactions of the Association for Computational Linguistics*, vol. 10, 2022, pp. 522–538.

[8] Artetxe, Mikel, et al. *On the Cross-Lingual Transferability of Monolingual Representations*. 2020.

[9] Kakwani, Divyanshu, et al. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-Trained Multilingual Language Models for Indian Languages. 2020.

[10] Khanuja, Simran, et al. MuRIL: Multilingual Representations for Indian Languages. 2021.

[11] Xue, Linting, et al. MT5: A Massively Multilingual Pre-Trained Text-To-Text Transformer. 2021.