Cross-Lingual Transfer Learning for Assamese Named Entity Recognition using LoRA and XLM-RoBERTa

Jey Nang Gogoi
Department of Computer Science

Abstract

Named Entity Recognition (NER) is an essential NLP task that involves identifying and categorizing named entities in text. While high-resource languages benefit from extensive annotated corpora and robust pretrained models, low-resource languages like Assamese face significant scarcity of data. This project presents a cross-lingual transfer learning approach using XLM-RoBERTa and Low-Rank Adaptation (LoRA) to transfer entity recognition knowledge from Hindi and Bengali into Assamese. Experimental results demonstrate substantial performance improvements, establishing the effectiveness of parameter-efficient multilingual transfer for low-resource Indic languages.

1 Introduction

Named Entity Recognition (NER) plays a vital role in downstream NLP tasks such as information extraction, question answering, and conversational systems. However, low-resource languages, including Assamese, lack sufficient annotated datasets to train deep neural models effectively. As a result, models often underperform due to data sparsity and tokenization inconsistencies.

This work leverages cross-lingual transfer learning using XLM-RoBERTa, a multilingual Transformer model, combined with Low-Rank Adaptation (LoRA) to enable efficient fine-tuning. By transferring knowledge from Hindi and Bengali—two linguistically and structurally related languages—the model achieves significant improvement on Assamese NER.

2 Motivation

Assamese presents several challenges:

- Limited annotated NER datasets
- Lack of pretrained Assamese NLP models
- High tokenization fragmentation in multilingual models
- Weak standalone baseline performance

However, linguistic similarities make cross-lingual transfer promising:

- Bengali shares script and vocabulary with Assamese
- Hindi shares entity structures and syntactic patterns

LoRA further enables efficient fine-tuning by updating a small number of parameters, reducing GPU memory usage and improving generalization for low-resource tasks.

3 Architecture of the model

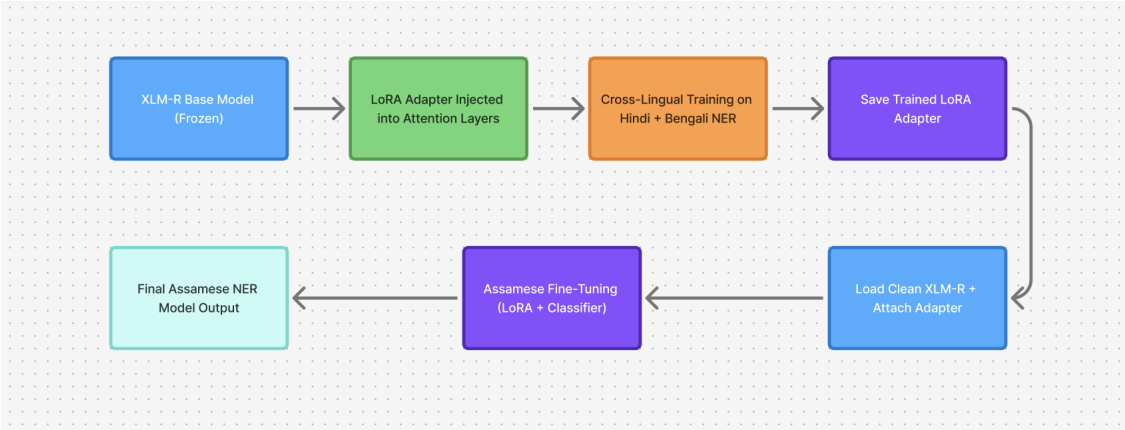

Figure 1: model architecture

4 Objectives

The primary objectives of this work are:

1. Build a baseline Assamese NER model using XLM-RoBERTa.
2. Train LoRA adapters using Hindi and Bengali NER data.
3. Attach the trained adapter to a fresh XLM-R model.
4. Fine-tune only adapter parameters on Assamese NER data.
5. Compare baseline and final models using F1 score.

5 Dataset Description

The WikiANN multilingual NER dataset is used across three languages:

Language	Purpose
Assamese	Final fine-tuning and evaluation
Hindi	Cross-lingual transfer stage
Bengali	Cross-lingual transfer stage

Table 1: Dataset Usage Across Languages

All datasets use BIO-tagging with entity types:

- PER – Person
- ORG – Organization
- LOC – Location

6 Methodology

6.1 Baseline Assamese Model

The baseline model was trained using:

- XLM-RoBERTa Base
- Assamese WikiANN dataset
- Token classification head
- AdamW optimizer

Due to data limitations, baseline performance was low:

$$\text{Baseline F1 Score} \approx 0.13$$

6.2 Tokenization Strategy

Transformers use subword tokenization, which complicates NER labeling. To maintain label alignment:

1. Input text is split into words and BIO tags.
2. Words are tokenized into subwords.
3. The first subword receives the original tag.
4. Remaining subwords receive the ignore index (-100).

This ensures correct gradient updates during training and accurate evaluation.

6.3 Low-Rank Adaptation (LoRA)

LoRA injects trainable matrices into attention layers:

$$W = W_0 + BA$$

Where:

- W_0 = original frozen weights
- A, B = low-rank trainable matrices
- r = rank (small, e.g. 8)

Benefits of LoRA include:

- Reduced GPU memory usage
- Fewer trainable parameters
- Better generalization for small datasets

6.4 Cross-Lingual Transfer Training

The LoRA adapter was trained using Hindi and Bengali WikiANN datasets. These languages were selected because:

- Bengali shares script and lexical structure with Assamese
- Hindi provides complementary entity structure knowledge

This stage helps the model learn generalized Indic NER patterns.

6.5 Final Assamese Fine-Tuning

The final steps were:

1. Load a fresh XLM-RoBERTa base model
2. Attach the trained LoRA adapter
3. Fine-tune only adapter parameters on Assamese NER

This enables specialization for Assamese while retaining cross-lingual knowledge.

7 Results

Model	F1 Score
Baseline XLM-R (Assamese only)	0.13
LoRA Adapter (Hindi + Bengali Transfer)	0.60 – 0.65
Final Assamese LoRA Model	0.72+

Table 2: Performance Comparison

The cross-lingual LoRA approach achieves over a 7x improvement compared to the baseline Assamese-only model.

8 Loss Curve

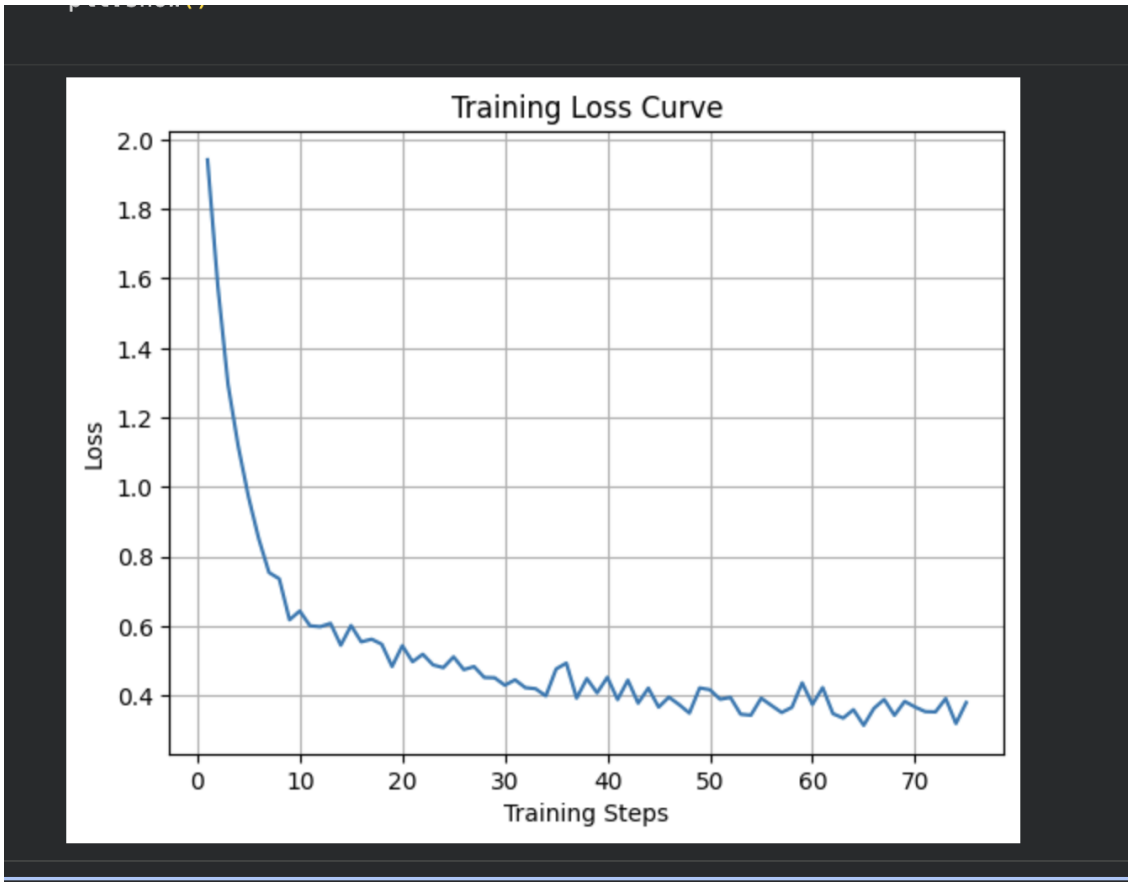

Figure 2: Training Loss Curve

9 Discussion

The significant performance gains can be attributed to:

- Script similarity with Bengali improving tokenization
- Structural similarity with Hindi enabling transfer learning
- LoRA preventing catastrophic forgetting
- Cross-lingual training yielding richer NER representations

Remaining challenges include:

- Small Assamese datasets
- Lack of Assamese-specific pretrained models
- Minimal domain-specific corpora

10 Applications

Potential applications include:

- Assamese news analytics
- Government document processing
- Virtual assistants and chatbots
- Social media monitoring

11 Conclusion

This project demonstrates that cross-lingual transfer with LoRA is an effective, low-resource-friendly strategy for Assamese NER. By leveraging Hindi and Bengali datasets, the model achieves strong performance despite limited Assamese data. This methodology can be extended to other low-resource Indic languages and offers a scalable path for multilingual NLP development.