Cross-Lingual Named Entity Recognition for Indian Languages

B.Tech Project

Rijul Singla (22075070)

Computer Science and Engineering, IIT (BHU)

Submitted to:

Prof. Anil Kumar Singh

Problem Statement

1

Task 1: Low-Resource NER

Develop Named Entity Recognition models for **Bhojpuri, Magahi, and Maithili** using transformer-based IndicNER architecture with heterogeneous datasets.

2

Task 2: Cross-Lingual Transfer

Generate high-quality NER datasets for **7 Indic languages** (Assamese, Bengali, Gujarati, Malayalam, Marathi, Tamil, Telugu) through cross-lingual projection from Hindi and train a multilingual model.

Task 1 Methodology: Low-Resource NER

Model Architecture

Base Model: **ai4bharat/IndicNER**

Task: Token-level classification for entity recognition

Training Datasets

- Naamapadam corpus
- BMM NER dataset
- HiNER annotations

Hyperparameters

- Learning Rate: **2e-5**
- Training Epochs: **10**
- Batch Size: **4–8**
- Optimizer: AdamW

Evaluation Metrics

Precision, Recall, F1-Score, and Overall Accuracy assessed for each language and entity type.

Task 2:
Cross-Lingual Dataset Generation Pipeline

Word Alignment

Apply SimAlign and Awesome-align algorithms to establish word-level correspondences between Hindi source and target language texts.

Tag Projection

Transfer NER tags from aligned Hindi entities to corresponding tokens in target languages while preserving entity boundaries.

Quality Validation

Validate projected annotations through entity matching verification, tag type consistency checks, and token-level alignment accuracy assessment.

Fine-tuned **XLM-RoBERTa** multilingual model with learning rates 2e-5 to 3e-5, 10 epochs, AdamW optimizer, and batch sizes 4–8 for optimal cross-lingual transfer.

Results: Low-Resource Language Performance

Bhojpuri

Precision	0.776
Recall	0.780
F1-Score	0.778
Accuracy	95.2%

Magahi

Precision	0.735
Recall	0.705
F1-Score	0.738
Accuracy	94.9%

Maithili

Precision	0.755
Recall	0.725
F1-Score	0.760
Accuracy	94.9%

Entity-Level Performance Analysis

Bhojpuri Entities

Entity	F1
PER	0.82
LOC	0.79
ORG	0.73

Magahi Entities

Entity	F1
PER	0.78
LOC	0.75
ORG	0.68

Maithili Entities

Entity	F1
PER	0.80
LOC	0.77
ORG	0.71

Person entities (PER) consistently achieved highest F1-scores across all three languages, while organization entities (ORG) proved most challenging to identify accurately.

Cross-Lingual Dataset Quality Metrics

75.7%

Entity Matching

Average accuracy across 7 target languages

91.6%

Tag Consistency

Entity type preservation rate

82%

Token Alignment

Word-level F1-score

Language-Wise Performance

Language	Entity Match	Tag Type	Token F1
Assamese	72.3%	89.5%	0.79
Bengali	78.1%	93.2%	0.84
Gujarati	76.8%	92.1%	0.83
Malayalam	73.5%	90.3%	0.80
Marathi	77.2%	92.8%	0.83
Tamil	74.9%	90.8%	0.81
Telugu	76.4%	91.5%	0.82

Multilingual Model Performance

XLM-RoBERTa fine-tuned on cross-lingually projected datasets for 7 Indic languages

Language	Precision	Recall	F1-Score	Accuracy
Assamese	0.762	0.748	0.755	97.2%
Bengali	0.812	0.798	0.805	98.3%
Gujarati	0.795	0.781	0.788	98.1%
Malayalam	0.771	0.759	0.765	97.5%
Marathi	0.803	0.789	0.796	98.2%
Tamil	0.778	0.765	0.771	97.7%
Telugu	0.786	0.773	0.779	97.9%
Aggregate	**0.787**	**0.773**	**0.789**	97.97%

Key Achievements

01

Low-Resource NER Success

Developed high-quality models for Bhojpuri, Magahi, and Maithili with F1-scores ranging from 0.738 to 0.778 and accuracies exceeding 94.9%.

02

Quality Dataset Generation

Created cross-lingual NER datasets for 7 Indic languages achieving 75.7% entity matching, 91.6% tag consistency, and 82% token-level F1.

03

Robust Multilingual Model

Trained XLM-RoBERTa achieving 78.91% aggregate F1-score and 97.97% accuracy across all 7 target languages.

04

Validated Transfer Approach

Demonstrated that alignment-based cross-lingual projection combined with multilingual transfer learning effectively addresses NER challenges in low-resource scenarios.

Conclusion

Research Impact

This work successfully demonstrates two complementary approaches to Named Entity Recognition for low-resource Indian languages.

The transformer-based models for Bhojpuri, Magahi, and Maithili establish strong baseline performance, while the cross-lingual projection methodology enables rapid dataset creation for seven additional languages.

The **78.91% F1-score** achieved by the multilingual model validates that alignment-based transfer learning is a viable strategy for expanding NER capabilities to underrepresented languages.

These results pave the way for improved natural language processing tools across the diverse linguistic landscape of India.