The Chitransh Academic & Research
International Journal of Multidisaciplinary Reseaech

An Online Peer Reviewed / Refereed Journal
Volume1 | Issue 04 |September 2025
ISSN: 3049-303X (Online)
Website: www.thechitranshacadmic.in

A Review of Cross-Lingual Transfer Learning Approaches for Low-Resource Named Entity Recognition

[1]Ms.Shruti Arya&[2]Dr. (Prof.) Vijay Kumar

[1]Phd Scholar, [2]Professor

[1,2]Jayoti Vidyapeeth Women's University, Jaipur

ARTICLEDETAILS

Research Paper

Received: **30/08/2025**

Accepted: **10/09/2025**

Published: **30/09/2025**

Keywords: NamedEntity Recognition, Low-Resource NLP, Cross-Lingual Transfer Learning, Multilingual BERT, Hindi, Marathi, Maithili, Haryanavi

ABSTRACT

Named Entity Recognition (NER) and slot filling are essential tasks in Natural Language Processing (NLP), enabling applications like chatbots, search engines, and automated customer support. However, low-resource languages (e.g., Hindi, Marathi, Thai) face challenges due to limited annotated data and insufficient pre-trained models. Cross-lingual transfer learning (CLTL) addresses this by leveraging high-resource languages (e.g., English) to improve performance in low-resource settings. This paper reviews three key studies on CLTL for NER and slot filling, comparing methodologies, datasets, and results. We highlight that multilingual embeddings (e.g., XLM-RoBERTa) outperform translation-based methods when small target-language data is available, while monolingual models (e.g., MahaRoBERTa) excel when language-specific pre-training is strong. We also discuss challenges, such as data sparsity and annotation inconsistencies, and suggest future directions, including hybrid embeddings and script-agnostic transfer learning.

INTRODUCTION-

Named Entity Recognition (NER) and slot filling represent two fundamental tasks in Natural Language Processing (NLP) that contribute significantly to the extraction of structured semantic information from unstructured text. NER is primarily concerned with the identification and categorization of entities such as persons, organizations, and locations, while slot filling facilitates the conversion of natural language utterances into structured Semantic frames, for instance, mapping the utterance 'Set an alarm for 7 AM' to the structured representation [action: set_alarm, time: 7 AM]. Considerable progress has been achieved for these tasks in high-resource languages, particularly English, where large-scale annotated datasets, robust pre-trained models, and supportive linguistic features (e.g., capitalization, standardized orthography) enable high levels of accuracy. In contrast, low-resource languages continue to pose significant challenges. The scarcity of annotated corpora, the limited availability of pre-trained models, and inherent linguistic complexities—such as the absence of capitalization in Indic languages like Hindi and Marathi and their rich morphological variation—constrain the direct applicability of existing approaches.

In response to these limitations, **Cross-Lingual Transfer Learning (CLTL)** has emerged as a promising paradigm, offering the potential to transfer knowledge acquired from high-resource languages to improve performance in low-resource settings. CLTL enables the development of models that leverage multilingual embeddings, shared subword vocabularies, and aligned semantic representations, thereby mitigating data scarcity and improving generalization across languages. Recent scholarship underscores the efficacy of this approach in advancing NER and slot filling for resource-constrained languages. For instance, Schuster et al. (2019) investigated CLTL for multilingual dialogue systems, demonstrating effective transfer of slot-filling capabilities from English to Spanish and Thai. Litake et al. (2023) conducted a comparative study of monolingual and multilingual BERT models for NER in Hindi and Marathi, revealing the advantages of multilingual pre-trained architectures in low-resource contexts. Furthermore, Sabane et al. (2023) introduced assisting-language strategies, wherein knowledge from linguistically related languages was exploited to enhance NER performance in under-resourced settings.

Building upon these contributions, this paper provides a systematic review of recent

Corresponding Author:Arya.shruti07@gmail.com

developments in CLTL for NER and slot filling, with a particular focus on applications in low-resource languages. The objective is to critically analyze methodological advancements, assess their effectiveness in diverse linguistic contexts, and highlight future research directions that can inform the design of more inclusive and equitable multilingual NLP systems

REVIEW OF LITERATURE

Low-resource language challenges have been increasingly tackled through research centered on cross-lingual transfer learning (CLTL) and Named Entity Recognition (NER), as well as in slot filling. The area received significant attention thanks to Schuster et al.'s (2019) creation of a multilingual dataset of annotated utterances in English, Spanish, and Thai, as well as their evaluation of transfer strategies such as translation-based transfer, cross-lingual embeddings, or multilanguage contextual encoders. In low-resource contexts, multilingual contextual representations are more effective than static embeddings and monolingual models can sometimes outperform cross-lingual methods when combined with limited target language data. With the release of BERT by Devlin, Chang, Leech, and Toutanova (2019) and its multilingual successor M-BERT from Pires, Schlinger & Garrette (2019), the field was further developed to enable deep bidirectional representation over 100 or more languages with good zero-shot performance even across different scripts. In spite of this, these models are more dependable for languages that share similar typologies and exhibit systematic flaws when dealing with highly structural or complex morphologies. Using both monolingual and multilingual transformer models, Litake (1923) by Sabane, Patil, Ranade, and Joshi utilized the Indian model in Hindi and Marathi NER to demonstrate that MahaRoBERTA is monolingua and not trilangular (Mural version) while XLM-Rebenerutheter has been found to be superior for Hindi. Joshi's work on the resource gap in 2022 was complemented by MahaCorpus and pre-trained Marathi models (MahaBERT, MahaRoBERTA, or MahaGPT), which showed significant progress in downstream tasks such as Sentiment analysis and NER, highlighting the critical role of specialized monolingual resources. Using related languages like Hindi, Sabane et al. (2023) demonstrated that strategically aligned multilingual training can achieve better results than randomly mixing the data in different ways, building on previous work. The research indicates that CLTL is a viable approach to managing data scarcity in low-resource settings, but its effectiveness depends

Corresponding Author:Arya.shruti07@gmail.com

on factors such as linguistic proximity, dataset quality, and the availability of monolingual resources. Therefore, it remains unclear how to balance generalization with language-specific specialization in multilingual NLP systems.

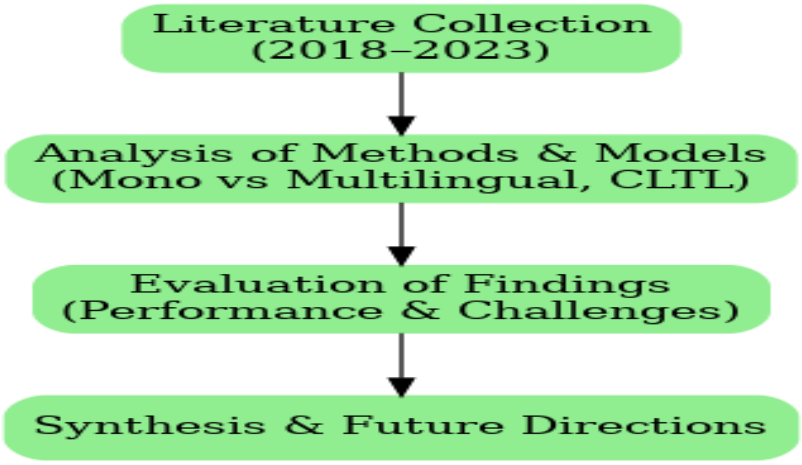
OBJECTIVE –

The objective of this research paper is to critically review and analyze recent advancements in cross-lingual transfer learning (CLTL) for Named Entity Recognition (NER) and slot filling, with a specific emphasis on applications in low-resource languages such as Hindi and Haryanvi , Maithili.

HYPOTHESIS –

H_0 -Cross-lingual transfer learning (CLTL) can enhance the performance of NER and slot filling in low-resource languages by leveraging high-resource languages. Monolingual models and assisting-language strategies are expected to further improve results when sufficient data or related languages are available.

RESEARCH METHODOLOGY



- Identification of relevant literature (2018–2023).
- Selection of studies focusing on CLTL, NER, slot filling, and low-resourcelanguages.
- Comparative analysis of methodologies, models, datasets, and strategies.
- Evaluation of results using reported metrics (F1-score, accuracy, etc.).
- Thematic synthesis of key challenges and emerging solutions.
- Identification of research gaps and future directions.

Corresponding Author:Arya.shruti07@gmail.com

Key Approaches and Findings

Cross-Lingual Transfer Methods

(A) Data Translation vs. Embedding Transfer (Schuster et al.)

- Training Data Translation: Translating English data to target languages (e.g., Spanish) works best in zero-shot settings (no target-language data).
- Multilingual Embeddings (e.g., CoVe, XLM-R): Perform better when small target-language data (100–200 examples) is available.
- Key Insight: Sharing BiLSTM/CRF layers across languages improves performance more than aligning word embeddings**.**

(B) Monolingual vs. Multilingual Models (Litake et al.)

- Marathi: Monolingual models (e.g., MahaRoBERTa) outperform multilingual ones (e.g., mBERT) due to specialized pre-training.
- Hindi: Multilingual models (e.g., XLM-RoBERTa) work better, indicating a need for improved Hindi-specific models.
- Cross-Language Testing: Marathi models generalize well to Hindi (shared Devanagari script), but Hindi models struggle with Marathi**.**

(C) Assisting-Language Strategies (Sabane et al.)

- Merging Hindi & Marathi Datasets: Improves NER performance for both languages.
- Challenge: Blindly merging all datasets can hurt performance; data selection (e.g., filtering divergent examples) is crucial.
- Best Model: XLM-RoBERTa performs well on merged datasets.

3. Datasets and Evaluation

Study	Languages	Datasets	Key Results
Schuster et al. (2019)	English, Spanish, Thai	57K task-oriented utterances	Multilingual embeddings > Translation when target data is limited
Litake et al. (2023)	Hindi, Marathi	IJCNLP, WikiAnn, IIT Bombay NER	MahaRoBERTa (Marathi) > XLM-R (Hindi)

Corresponding Author:Arya.shruti07@gmail.com

Study	Languages	Datasets	Key Results
Sabane et al. (2023)	Hindi, Marathi	Merged datasets	XLM-RoBERTa best for cross-lingual NER

CHALLENGE

Despite significant advancements in cross-lingual transfer learning (CLTL), several challenges persist in the development of robust Named Entity Recognition (NER) and slot filling systems for low-resource languages. A major limitation is the lack of annotated datasets, which restricts the effectiveness of supervised learning approaches. This scarcity is compounded by the limited availability of high-quality pre-trained models for underrepresented languages, in contrast to the extensive resources available for English and other high-resource languages. Furthermore, the performance of multilingual models such as M-BERT and XLM-R is uneven, as cross-lingual transfer tends to be less effective for typologically distant or morphologically rich languages, where structural divergences hinder accurate knowledge transfer. Indic languages such as Hindi and Marathi illustrate these difficulties, as they exhibit linguistic complexities including absence of capitalization, high levels of inflectional morphology, and orthographic variation, all of which reduce the accuracy of entity recognition.

Another challenge arises from script diversity, since languages employing scripts like Devanagari face difficulties in ensuring uniform representation and transfer across multilingual models. Moreover, the domain mismatch between training data (e.g., Wikipedia, news) and real-world applications (e.g., conversational or code-mixed data) further limits the applicability of existing models. Data imbalance across languages also undermines multilingual training, as high-resource languages dominate and overshadow the representation of low-resource counterparts. While assisting-language strategies—leveraging related languages such as Hindi to support Marathi—show promise, they also carry the risk of negative transfer if resources are not carefully aligned. Finally, there remains a critical trade-off between generalization and specialization: multilingual models demonstrate strong cross-lingual generalization, but often fail to capture the fine-grained linguistic nuances that specialized monolingual models can provide. These challenges underscore the need for more inclusive approaches that balance multilingual

Corresponding Author:Arya.shruti07@gmail.com

transfer with language-specific adaptation.

FUTURE WORK

Future research on cross-lingual transfer learning (CLTL) for Named Entity Recognition (NER) and slot filling should focus on overcoming the persistent barriers faced by low-resource languages. A key priority is the development of larger and more diverse annotated corpora, particularly domain-specific datasets that better reflect real-world applications. Parallelly, the expansion of monolingual pre-trained models for languages such as Hindi and Marathi is crucial, as these models have shown superior adaptability to language-specific morphological and syntactic patterns compared to general-purpose multilingual models. Enhancing multilingual architectures through balanced training across languages and better script coverage also remains an important avenue to improve their cross-lingual effectiveness.

Another promising direction lies in refining assisting-language strategies, where related languages are leveraged in a structured and selective manner to maximize positive transfer while minimizing negative effects. Techniques such as language-adaptive pre-training, selective fine-tuning, and adversarial learning could further strengthen transfer performance. Addressing domain adaptation challenges is equally essential, given the increasing prevalence of conversational, noisy, and code-mixed data in practical settings. Finally, long-term efforts should aim at building inclusive multilingual NLP frameworks that extend beyond high- and mid-resource languages to incorporate truly low-resource and endangered languages, thereby reducing linguistic inequalities in digital technologies

REFERENCES

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep

bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186. Association for Computational Linguistics.

2. Joshi, R. (2022). L3Cube-MahaCorpus and MahaBERT: Marathi monolingual corpus, Marathi BERT language models, and resources. Proceedings of the WILDRE-6 Workshop @ LREC, 97–101.
3. Litake, O., Sabane, M., Patil, P., Ranade, A., & Joshi, R. (2023). Mono versus multilingual BERT: A case study in Hindi and Marathi named entity recognition. In Advances in Intelligent Systems and Computing (pp. xxx–xxx). Springer.
4. Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? Proceedings of ACL, 4996–5001.
5. Sabane, M., Ranade, A., Litake, O., Patil, P., Joshi, R., & Kadam, D. (2023). Enhancing low-resource NER using assisting language and transfer learning. arXiv preprint arXiv:2306.06477.
6. Schuster, S., Gupta, S., Shah, R., & Lewis, M. (2019). Cross-lingual transfer learning for multilingual task-oriented dialog.
7. Arora, G. (2020). inltk: Natural language toolkit forindic languages. arXiv preprint arXiv:2009.12534.
8. Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T.(2017). Enriching word vectors with sub word information.Transactions of the Association for Computational Linguistics, 5:135–146.
9. Corneau, A., Khandelwal, K., Goyal, N., Chaudhary,V., Wenzek, G., Guzm´an, F., Grave, E., Ott, M.,Zettlemoyer, L., and Stoyanov, V. (2019). Unsupervisedcross-lingual representation learning at scale.
10. Delobelle, P., Winters, T., and Berendt, B. (2020).Robbert: a dutch roBERTa-based language model.
11. Patil, O., Sabane, M., & Joshi, R. (2023). Cross-Lingual Transfer Between Related Indian Languages: A Case Study of Hindi and Marathi.
12. Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N., Bhattacharyya, A., Khapra, M. M., & Kumar, P. (2020). IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks, and Pre-trained Multilingual Language Models for Indian Languages.

13. Pires, T., Schlinger, E., & Garrette, D. (2019). How Multilingual is Multilingual BERT?
14. Dandapat, S., Sarkar, S., & Basu, A. (2022). L3Cube-MahaNER: A Marathi Named Entity Recognition Dataset and BERT Models. LREC.
15. Singh, T. D., Ekbal, A., & Saha, S. (2021). Multi-Task Learning for Named Entity Recognition in Indian Languages.