INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	Deepen Naorem
Roll Number	:	186155102
Programme of Study	:	Ph.D.

Thesis Title: Cross-Lingual Embedding between Non-Isomorphic Language Pairs (Special Focus on English and Manipuri)

Name of Thesis Supervisor(s)	:	Prof. Sanasam Ranbir Singh & Prof. Priyankoo Sarmah
Thesis Submitted to the Academic Division	:	Centre for Linguistic Science and Technology
Date of completion of Thesis Viva-Voce Exam	:	10/10/2025
Key words for description of Thesis Work	:	CLWE, BDI, Embeddings, Morphology, and Non-Isomorphic

---

SHORT ABSTRACT

Research in Natural Language Processing (NLP) has primarily focused on resource-rich languages like English, leaving low-resource languages underrepresented and contributing to a phenomenon known as the digital divide. This disparity limits the development of NLP tools for low-resource languages, such as Manipuri, a morphologically rich Tibeto-Burman language. Transfer learning, leveraging resource-rich languages, has emerged as a solution to this challenge, with cross-lingual embeddings playing a pivotal role in aligning lexical units between languages. This thesis first presents a comprehensive empirical evaluation of cross-lingual embeddings between English and Manipuri, distant language pairs, in BDI using state-of-the-art supervised and unsupervised approaches. The findings highlight that the non-isomorphic nature of the language pairs degrades the cross-lingual embedding quality, making dictionary pair selection crucial, and the morphological richness of the target language further impacts BDI performance. This thesis proposes two novel approaches to address the challenges posed by structural and morphological disparities in distant language pairs. First, a ridge regression-based orthogonal mapping method is introduced, incorporating graph centrality for improved dictionary alignment, outperforming conventional orthogonal mapping techniques, particularly for structurally distant languages like English-Manipuri. Second, a contrastive learning-based method is developed to leverage the morphological richness of Manipuri. Experimental results across several language pairs show significant improvements in BDI, machine translation, and cross-lingual sentence retrieval tasks, outperforming baseline methods. Furthermore, with the increasing advancement of Large Language Models (LLMs), this thesis evaluates the performance of unsupervised, supervised, and few-shot prompting approaches using large language models (LLMs) for BDI across distant language pairs. The findings reveal that few-shot prompting, leveraging minimal examples, consistently outperforms unsupervised and supervised methods, demonstrating robustness against over-fitting and cost-effectiveness for low-resource languages. These results suggest that few-shot prompting is a powerful alternative for multilingual BDI tasks, with future work focusing on prompt optimization.