INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

Cross-Lingual Embedding between Non-Isomorphic Language Pairs

(Special Focus on English and Manipuri)

by

Deepen Naorem

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the

CENTRE FOR LINGUISTIC SCIENCE AND TECHNOLOGY

Under the supervision of

Prof. Sanasam Ranbir Singh and Prof. Priyankoo Sarmah

October 2025

Declaration of Authorship

I, Deepen Naorem, hereby declare that:

- The research work presented in this thesis is my original work, conducted under the general guidance of my supervisors.
- This thesis has not been submitted to any other institution for the award of a degree or diploma.
- Whenever I have incorporated materials (data, theoretical analysis, or results) from external sources, appropriate credit has been given by citing the original authors/researchers within the thesis and listing their details in the references.
- Any direct quotations from external works have been properly attributed to their respective sources.

Deepen Naorem
Research Scholar,
CENTRE FOR LINGUISTIC SCIENCE AND TECHNOLOGY,
Indian Institute of Technology Guwahati,
Guwahati, Assam, INDIA 781039,
deepennaorem@iitg.ac.in, deepennaorem@gmail.com

Place: IIT Guwahati

Certificate

This is to certify that the thesis entitled **"Cross-Lingual Embedding between Non-Isomorphic Language Pairs (Special Focus on English and Manipuri)"** being submitted by **Mr. Deepen Naorem** to the *Centre for Linguistic Science and Technology, Indian Institute of Technology Guwahati*, is a record of bonafide research work under my supervision and is worthy of consideration for the award of the degree of Doctor of Philosophy of the Institute.

Prof. Sanasam Ranbir Singh
Department of CSE,
Indian Institute of Technology Guwahati,
Guwahati, Assam, INDIA 781039,
ranbir@iitg.ac.in

Prof. Priyankoo Sarmah
Department of Humanities and Social Science,
Indian Institute of Technology Guwahati,
Guwahati, Assam, INDIA 781039,
priyankoo@iitg.ac.in

Place: IIT Guwahati

Dedicated to

my father Late Naorem Dhiren

and

my mother Soibam Premi Devi

& all of my teachers and gurus

for their infinite love, support, motivation and guidance.

Acknowledgements

I am deeply grateful to everyone who directly or indirectly contributed to the successful completion of my Ph.D. journey. First and foremost, I extend my heartfelt thanks to my supervisors, Prof. Sanasam Ranbir Singh and Prof. Priyankoo Sarmah, for their exceptional and inspiring guidance throughout this journey. Their unwavering dedication to their responsibilities and research has been a constant source of motivation for me. I will always remain indebted to them for their invaluable insights, stimulating research discussions, and for instilling in me a strong sense of professional work ethics.

I would like to express my gratitude to my Doctoral Committee members, Prof. Sukumar Nandi, Prof. Ashish Anand, and Prof. Bidisha Som, for their invaluable suggestions in shaping both my research objectives and my thesis as a whole. I am also deeply thankful to Dr. Ashish Anand for our insightful research discussions and his valuable guidance. Additionally, I extend my sincere appreciation to Prof. Sanasam Ranbir Singh, Head of the Centre for Linguistic Science and Technology, along with other faculty members, for their direct and indirect support throughout my PhD journey.

I sincerely express my gratitude to Mr. Souvik Chowdhury, Mr. Santanu Majumdar, Mr. Raktajit Pathak, Mr. Nanu Alan Kachari, Mr. Bhriguraj Borah, and all the institute's staff for their invaluable support in making my journey smooth and productive. I am especially grateful to Mr. Santanu Majumdar for his unwavering dedication in managing efficient computing facilities at the Centre, without which my thesis would not have been completed on time. Additionally, I extend my thanks to the IIT Guwahati administration for providing on-campus hostel accommodation. From the bottom of my heart, I deeply appreciate the efforts of the administrative staff from the Students Affairs Section, Academic Section, Research and Development Section, hostel caretakers, mess staff, canteen staff, security personnel, and housekeeping staff for ensuring a comfortable and memorable stay.

Having good friends is truly a blessing, and I am fortunate to have a large circle of close and supportive friends with whom I have shared many meaningful moments. I feel privileged to mention Rokesh Laishram, Dr. Thockchom Birjit Singha, Rakesh Singha, Jitender Ngangom, and Dr. Debojit Paul as longtime peers and colleagues who have supported me in various aspects throughout my journey.

I have also been fortunate to have incredibly helpful and supportive seniors, including Dr. Durgesh Kumar, Rajib Chakrabartty, Dr. Akash Anil, Dr. Anasua Mitra, Dr. Pradeepkumar Bhale, Dr. Dhrubajyoti Pathak, Dr. Sanasam Sunderlal, and Dr. Franco Mayanglambam, whose guidance has been invaluable.

My time at IIT Guwahati was made even more memorable through the wonderful experiences shared with members of the OSINT Lab, such as Akash Anil, Neelakshi Sarma, Hemanta Baruah, Loitongbam Gyanendro Singh, Roshan Singh, Lenin Laitonjam, Pankaj Choudhury, Tonmoya Sarmah, Anasua Mitra, Rajib Chakrabartty, Soumyadeep Jana, Saurabh Kumar, Mridul Jyoti Roy, Tarik Mohammad Saikia, Okram Jimmy Singh, Jibon Kumar Borgoyary, and many more. Their presence enriched my journey, making it all the more special.

I would also like to express my gratitude to the members of the CLST lab, including Mr. Telem Joyson Singh, Mrs. Maisang Kamei Salice, Mr. Hemanta Baruah, Mrs. Emily Thomas, Mr. Chaitanya Kirti, Mrs. Joyshree Chakraborty, Mr. Pankaj Choudhury, Mrs. Meghali Deka, and many others.

My family has been a pillar of strength throughout my academic journey, providing unwavering support, love, motivation, and faith that have helped me overcome challenges at various stages of my life. From the bottom of my heart, I extend my deepest gratitude to my mother, Soibam Premi Devi, and my brother, Dr. Deepak Naorem, for their immense love and moral support. Lastly, I sincerely thank my friends, family, relatives, and everyone who has contributed to my academic growth and success.

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. Many of these languages lack parallel corpora, essential for various NLP tasks. 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 alignment is foundational for cross-lingual downstream NLP tasks like Bilingual Dictionary Induction (BDI) and Machine Translation. 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. Experimental results across multiple language pairs demonstrate its efficacy in BDI tasks. Second, a contrastive learning-based method is developed to leverage the morphological richness of Manipuri. Unlike traditional methods of morphological segmentation, this approach utilizes the relationship between roots and affixes to enhance cross-lingual embedding quality. 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 overfitting 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.

Contents

List of Figures

List of Tables

Abbreviations

NLP	Natural Language Processing
CLWE	Cross-lingual Word Embedding
NER	Named Entity Recognition
SOV	Subject-Object-Verb
SVO	Subject-Verb-Object
En-It	English-Italian
En-Mn	English-Manipuri
En-Fi	English-Finish
En-Tr	English-Turkish
En-Hi	English-Hindi
En-Ja	English-Japanese
En-Ta	English-Tamil
SVD	Singular Value Decomposition
BDI	Bilingual Dictionary Induction
LMs	Language Models
LLMs	Large Language Models
MACE	Morphology-Aware Cross-Lingual Embeddin
SOTA	State of the Art
MT	Machine Translation
CSRT	Cross-lingual Sentence Retrieval Task
CCA	Canonical Correlation Analysis
POS	Part-of-Speech
CBOW	Continuous Bag of Words
MUSE	Multilingual Unsupervised and Supervised Embedding
CSLS	Cross-domain similarity local scaling
MBERT	a Multilingual-BERT
BERT	Bidirectional Encoder Representations from Transformers
P@1	Precision at 1
P@5	Precision at 5

EMBERT	English-Manipuri BERT
GANs	Generative Adversarial Networks
FA	Frequency Align
DCA	Degree Centrality Align
DCAR	Degree Centrality Align and Regularised
ECA	Eigen Centrality Align
ECAR	Eigenvector Centrality Align and Regularised
FACR	Frequency Align Centrality Regularised
FAFR	Frequency Align Frequency Regularised
CA	Centrality Align
CACR	Centrality Align Centrality Regularised
CAFR	Centrality Align Frequency Regularised

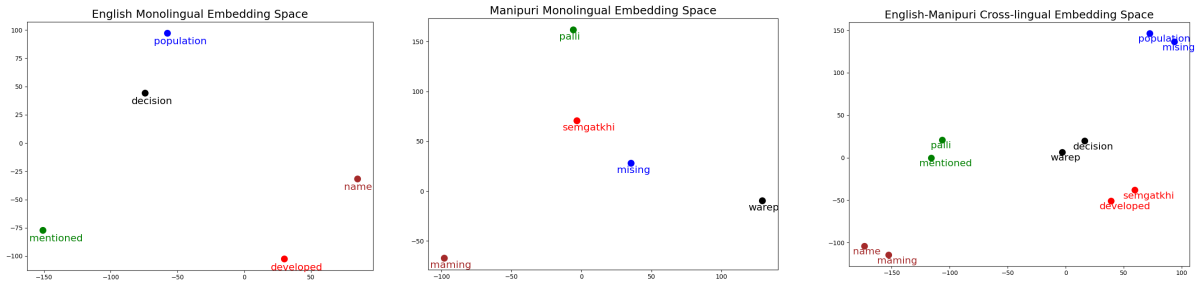
Chapter 1

Introduction

Word embedding [4, 5] is a low-dimensional dense vector representation of words, which captures semantic relations between words. Word embeddings are generated based on the distributional hypothesis that *words are known by the company they keep* [6]. Today, word embedding is pervasive and essential for developing various language processing, understanding, and natural language processing (NLP) tools. Most of the earlier studies on word embedding focus on monolingual setups, i.e., it considers text corpus in a given language and generates word embedding of the words in the corpus. English is the dominant language on digital platforms, so earlier studies on word embedding mostly focused on the English language. With the increase of non-English language content over digital media, especially on the Internet, researchers have also started to pay attention to the word embedding of non-English languages.

With growing demands for applications such as Bilingual Dictionary Induction, Machine Translation, and the increasing presence of multilingual content on the Internet, the need to share words of similar semantic meaning or transfer knowledge across languages also increases. An approach to handle this challenge is to devise methods for *cross-lingual word embedding* (CLWE). *Given two monolingual embeddings generated from two monolingual corpora, the task of cross-lingual word embedding is to project embedding vectors of words in one corpus (source language) to the embedding space of another corpus (target language) such that words with similar semantic meaning across source and target languages are close to each other.* Figure 1.1 shows an illustrative example of cross-lingual word embedding, where words in two monolingual corpora in English and Manipuri are mapped to a shared space.[1] From the plots, it is evident that the words with similar semantic meanings (represented by the same color) can map closer to each other in the shared space using a CLWE method.

[1]This mapping has been generated using VecMap[3], one of the popular CLWE methods.



(a) (b) (c)

Figure 1.1: From Monolingual (a) English and (b) Manipuri Embedding Space to (c) Shared space between English and Manipuri language through Cross-lingual word embedding.

The study of CLWE is important from various perspectives. Some of them may be noted below.

- **Applications:** Monolingual embeddings are not feasible across languages and cannot be used in multilingual or cross-lingual NLP downstream applications, like Bilingual Dictionary Induction and Machine Translation[7]. Monolingual embeddings are generated by training a neural network separately for a particular language. Hence, words in different languages are mapped to different vector spaces. Since the embeddings are trained independently, the spatial distribution of words in one language does not correspond to that of another, thereby creating a misalignment when applied to multilingual or cross-lingual NLP tasks. Cross-lingual embeddings bring two monolingual embedding spaces into a shared space where words with similar semantic meanings across languages are close, facilitating cross-lingual NLP tasks like BDI and Machine Translation.

- **Multilingual contents:** Most of the earlier studies on word embedding have primarily focused on monolingual English setups. The need to generate embedding for non-English languages arises with the increase in digital media's multilingual and non-English text content. Based on the Statista report [2], close to 48% of the Internet content is in non-English languages. Cross-lingual embeddings help address multilingual content by mapping words from different languages into a shared vector space, ensuring that words with similar meanings have similar representations regardless of the language.

- **Resource-poor languages:** Though the development of monolingual word embedding for non-English languages is important, due to disparities in the availability of

[2] https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/

linguistic resources and corpus across languages, developing word embeddings for different languages with equivalent capabilities in downstream NLP tasks may not be feasible. In addition, the development of monolingual word embeddings for different languages separately may have limitations in knowledge sharing between languages. The disparity in linguistic resources and corpus across languages has a more significant impact, often referred to as the Digital Language Divide[3], highlights the gap between languages in terms of their digital presence, resources, and accessibility on the internet and digital platforms. This divide affects how different language communities access information, communicate, and participate in the digital world, thereby influencing human connections online and offline, as noted by Tim Berners-Lee: *The Web does not just connect machines; it connects people.* To address this divide, researchers have adopted transfer learning, leveraging resource-rich languages like English to build tools for resource-constrained languages. An important breakthrough in leveraging resource-rich language was the development of Cross-lingual Word Embeddings (CLWE)[8].

This thesis focuses on developing cross-lingual embedding between English and Manipuri language pairs. Manipuri, also known as Meiteilon, is a Tibeto-Burman language predominantly spoken in the northeastern Indian state of Manipur. Manipuri is a morphologically rich language with a strong presence of derivational morphology, where affixes are used to form new words from root words[9]. Manipuri language morphological system is more complex, often involving multiple affixations in a single word. Unlike an English sentence, which follows an SVO (Subject-Verb-Object) structure, the Manipuri sentence follows an SOV (Subject-Object-Verb) word order[10]. Such linguistic characteristics of the Manipuri language pose challenges in cross-lingual embedding with languages like English due to its nonisomorphic structure with the English language. While CLWE methods between isomorphic language pairs like English-Italian, English-Spanish, and Italian-French, which are linguistically close, have been extensively studied and shown encouraging performances[11], there are challenges when the language pairs are linguistically distant[12]. Existing CLWE methods perform poorly on nonisomorphic language pairs. The performance further degrades when one of the languages in the pair is morphologically rich language [13, 11, 14].

A widely adopted approach in cross-lingual is the mapping-based method[8], where word embeddings from different languages are transformed into a shared vector space, facilitating the alignment of lexical resources. Among the mapping-based methods, matrix factorization techniques, such as VecMap[3], have demonstrated strong performance for isomorphic or similar language pair, such as English-Italian (En-It). The same mapping-based methods often fail for distant language pairs such as English-Manipuri[15]. Manipuri language

[3] http://labs.theguardian.com/digital-language-divide/

presents unique challenges for CLWE due to its complex morphological structure [16] and distinct sentence structure [17]. Morphological-rich language causes inaccurate alignment in the shared embedding space and hampers performance in the BDI tasks. For an English word, the nearest 5 Manipuri words obtained with CLWE are inflected with suffixes, slightly changing the meaning compared to the ground truth Manipuri words, degrading the performance in BDI tasks. To further boost the BDI performance, contrastive learning cross-lingual embedding approach[1] has emerged as a powerful approach that combines static word embeddings (VecMap) with contextual embeddings generated by language models (LMs) and large language models (LLMs). The contrastive learning CLWE offers a promising result for isomorphic language pairs like English-Italian but fails to perform well in English-Finnish[1]. Like the Manipuri language, Finnish is a morphologically rich language.

This thesis investigates CLWE for the English-Manipuri language pair to understand the challenges mentioned above in greater detail, focusing on supervised and unsupervised approaches. Due to the lack of parallel corpora, the feasibility of comparable corpora, which describe the same events in different languages, is explored. The thesis evaluates the efficacy of CLWE methods for structurally distant language pairs, English-Manipuri, and identifies challenges, such as the non-isomorphic nature of English-Manipuri language pairs. Key findings indicate that traditional frequency-based seed dictionary selection methods are unsuitable for non-isomorphic language pairs like English-Manipuri. Instead, network centrality measures offer a more reliable criterion, as they capture the structural properties of word co-occurrence networks. This thesis proposes a graph centrality-aware ridge regression-based orthogonal mapping method that penalizes less comparable dictionary pairs and enhances the CLWE between non-isomorphic language pairs. Experimental results demonstrate improved performance in tasks such as bilingual dictionary induction (BDI), cross-lingual sentence retrieval, and machine translation. Furthermore, rich morphological features in Manipuri degrade the quality of embeddings due to inaccurate mappings of complex words. To tackle the challenges attributed to the morphologically rich nature of the target language, a novel Morphology-Aware Cross-Lingual Embedding (MACE) framework is introduced, leveraging contrastive learning to bring words with the same root closer while pushing unrelated words apart. Experiments reveal that this approach significantly improves BDI performance and other downstream applications compared to baseline methods.

With the success of the Large Language Model (LLM) and subsequent prompting method in many NLP tasks, studies were made to perform BDI using a few-shot prompting approach to leverage LLM[18]. However, this study neglects distant language pairs and language pairs where one of the languages in the pair is morphologically rich. This thesis evaluates the efficacy of few-shot prompting and fully supervised fine-tuning of LLMs for BDI tasks.

The thesis provides a comparative analysis of their strengths and limitations across linguistically distant language pairs. While few-shot prompting leverages in-context learning capabilities [19] with minimal supervision, fully supervised methods require extensive labeled datasets but offer higher accuracy. This thesis systematically examines their trade-offs for the first time, focusing on challenging language pairs like English-Manipuri, English-Tamil, and English-Japanese.

1.1 Challenges

While CLWE methods have been extensively studied for isomorphic language pairs, there are a few generic challenges, as listed below:

- **Non-Isomorphic nature of distant languages pairs**: CLWE methods perform poorly for non-European due to linguistic differences and the non-isomorphic nature of their embedding spaces [13, 11, 14]. For example, mapping-based methods often fail for distant language pairs such as English-Manipuri or English-Tamil.
- **Morphological Rich Nature of Manipuri Language**: Manipuri is a low-resource language and presents a unique challenge for CLWE due to its complex morphological structure [16]. Morphological-rich language causes inaccurate alignment in the shared embedding space.
- **Different Sentence Structure**: English and Manipuri exhibit different word orders in their syntax, with English following the Subject-Verb-Object (SVO) structure and Manipuri following the Subject-Object-Verb (SOV) structure. This fundamental syntactic difference has significant implications for natural language processing (NLP), particularly in machine translation, cross-lingual embeddings, and syntactic parsing.
- **Data sparsity**: The Manipuri language, like many other languages, incorporates loanwords—words borrowed from other languages due to historical, cultural, or social exchanges. While this enriches the language, it also provides challenges to natural language processing (NLP). Loanwords are often less common than native vocabulary in a Manipuri monolingual corpus. The infrequent occurrence of these words leads to data sparsity—an issue where certain words have limited or no representation in the training data, making it hard to model their semantic or syntactic properties accurately.
- **LLMs are underexplored for distant language pairs**: Recent research has explored the potential of large language models (LLMs) to induce BDI using a few-shot

prompting approach[18]. However, these studies have primarily focused on a limited set of resource-rich or structurally similar language pairs, often neglecting the evaluation of linguistically distant and resource-poor pairs such as English-Manipuri (En-Mn), English-Finnish (En-Fi), English-Turkish (En-Tr), English-Hindi (En-Hi), English-Japanese (En-Ja), and English-Tamil (En-Ta). Moreover, specific evaluation in LLM models like LLaMA remains unexplored, mainly in these challenging settings.

1.2 Objectives and Scope of the Thesis

To address some of the main challenges mentioned above, this thesis aim to address the following research objectives:

- **Empirical Study on English-Manipuri CLWE**: This study aims to observe the performance of various SOTA methods on CLWE between English and Manipuri. This evaluation aims to highlight significant challenges in English-Manipuri CLWE in greater detail.
- **Limitation of Orthogonal Mapping in non-isomorphic language pairs**: This objective demonstrates the limitations of frequency-based seed dictionary selection used in orthogonal matrix factorization CLWE method for non-isomorphic language pairs.
- **CLWE in Morphological rich language pairs**: This objective highlights the adverse effect on CLWE in BDI tasks when the target language in the language pair is morphologically rich.
- **Efficacy of LLMs for Linguistically distant Language Pairs in BDI**: The objective of this study is to understand the effectiveness of LLM for linguistically distant language pairs in BDI tasks.

1.3 Contributions Made in the Thesis

Based on the above-outlined objectives that tackle key challenges in CLWE between English and Manipuri, the thesis made the following contributions:

1.3.1 Empirical evaluation of English-Manipuri CLWE

As pointed out in the first objective, to understand the significant challenges in English-Manipuri CLWE in detail, this thesis empirically examines the efficacy of various state-of-the-art cross-lingual embedding methods for the English-Manipuri language pair [9]. The majority of cross-lingual embedding methods discussed in existing literature [8, 20, 21, 3, 22, 23, 24, 25] are predominantly designed for Isomorphic language pairs like English-Spanish, English-Italian, French-German, which share closer etymological ties [13, 11, 14]. Further empirical analysis is done to understand the linguistic factors affecting cross-lingual embedding quality. Experimental results indicate that cross-lingual embeddings perform better for linguistically similar language pairs, such as English-Italian than English-Manipuri. Based on the experimental findings and error analysis, it is evident that the linguistic differences like unbalanced frequency between source word and target word in the English-Manipuri bilingual dictionary pose challenges for achieving high-quality cross-lingual embeddings, mainly when evaluated in BDI tasks as compared to that of English-Italian. The sparse nature of Manipuri data, resulting from its complex morphological structure, exacerbates these challenges. It is evident that commonly used techniques, such as mapping, exhibit limited efficacy when applied to distant language pairs. This limitation may stem from the non-isomorphic nature and linguistic disparities, including morphological differences between the source and target languages.

1.3.2 Graph Centrality aware CLWE

As mentioned in the second objective, while projection-based CLWE methods demonstrate strong performance for isomorphic language pairs, as noted in [12], they tend to perform poorly for non-isomorphic language pairs. All previously reported orthogonal mapping approaches [26, 3, 23, 12] have relied on selecting seed bilingual dictionary pairs based on the most frequent words in the source language. This selection assumes that word frequencies in the source and target monolingual corpora are comparable, which does not hold for non-isomorphic language pairs. Addressing these limitations in frequency-based dictionary

alignment, this thesis introduces a *graph centrality-aware ridge regression-based orthogonal mapping method*, which: (i) improves the rank correlation of source and target dictionary word pairs across their respective monolingual corpora by leveraging graph centrality, and (ii) mitigates the impact of less comparable (misaligned) word pairs when estimating the projection matrix using ridge regression. Extensive experiments conducted across six state-of-the-art orthogonal frameworks on five different language pairs—both isomorphic and non-isomorphic—demonstrate that the proposed centrality-aware ridge regression method enhances performance in multiple applications, namely *BDI, Cross-lingual Sentence Retrieval Task (CSRT), and Machine Translation (MT)*. Additionally, an ablation study validates the effectiveness of centrality-based ridge regression.

1.3.3 Morphology Aware CLWE

As pointed out in the third objective, morphological-rich language causes inaccurate mappings while handling complex words because the existing CLWE models do not exploit sub-word level morphological information while aligning complex words [27]. It is observed that segmenting morpheme (root) and suffixes help in improving BDI performance [27, 28]. However, effective stemmers are only available for some languages, especially rich-resource ones. Therefore, we need CLWE models to enhance BDI performance for morphological-rich languages without root and suffix segmentation. However, so far, CLWE embedding methods such as VecMap[23] and [1] fail to handle the inaccurate mapping due to complex morphological properties in the BDI task. This CLWE characteristic hinders the performance of BDI. Motivated by the above concerns, this thesis proposes MACE (Morphology Aware Cross-lingual Embedding using Contrastive Learning), which brings target language words with the same root closer and pushes target words with different roots apart using *contrastive learning*. The experimental observation shows that bringing target words with the same root closer and target words with different roots apart using contrastive learning improves the BDI task's performance. Experiments on Machine Translation and Cross-lingual Sentence Retrieval Tasks also show that the proposed method outperforms the baseline method.

1.3.4 Evaluation of LLMs in Linguistically distant Language pairs

As mentioned in the fourth objective, recent research has explored the potential of large language models (LLMs) to induce BDI using a few-shot prompting approach[18]. This technique involves leveraging the inherent contextual understanding of LLMs to align embeddings with minimal supervision[29], often achieving better results than traditional VecMap

and contrastive learning methods. However, the above-mentioned studies have primarily focused on a limited set of resource-rich or structurally similar language pairs, often neglecting the evaluation of linguistically distant and resource-poor pairs such as English-Manipuri (En-Mn), English-Finnish (En-Fi), English-Turkish (En-Tr), English-Hindi (En-Hi), English-Japanese (En-Ja), and English-Tamil (En-Ta). Moreover, specific evaluation in LLM models like LLaMA still needs to be explored in these challenging settings. This thesis systematically evaluates unsupervised, supervised, and few-shot prompting in BDI tasks, analyzing their performance across varied linguistic language pairs. Specifically, the thesis assesses the efficacy of few-shot prompting and supervised fine-tuning in translation accuracy and adaptability to low-resource settings. The results show that all LLMs consistently perform better on linguistically similar pairs (e.g., En-It) than distant language pairs and pairs with morphologically complex language (e.g., En-Mn, En-Fi, En-Ta). The 5-shot prompting approach outperforms unsupervised and zero-shot settings in all cases and even surpasses supervised settings in 82.86% of cases. The findings suggest few-shot prompting as a cost-effective and powerful alternative for BDI, with future work focusing on prompt optimization.

1.4 Organization of the Thesis

The remaining part of the thesis is organized as follows:

- **Chapter 2: Literature Survey**: This chapter thoroughly reviews CLWE from a mapping-based approach to a state-of-the-art few-shot prompting approach in BDI.
- **Chapter 3: Empirical Evaluation on English-Manipuri CLWE**: This chapter presents a comprehensive empirical evaluation of cross-lingual embeddings between English and Manipuri for bilingual dictionary induction (BDI) using supervised and unsupervised approaches.
- **Chapter 4: Improving Linear Orthogonal Mapping approach** : This chapter demonstrates the limitations of frequency-based seed dictionary selection for non-isomorphic language pairs and proposes graph-centrality-aware ridge regression for improved CLWE performance.
- **Chapter 5: A Novel Morphology Aware CLWE framework**: This chapter presents a contrastive learning framework to improve CLWE for morphologically rich languages without explicit root and suffix segmentation.
- **Chapter 6: Evaluation SOTA LLMs for Linguistically distant Language pairs**: This chapter systematically evaluates few-shot prompting and fully supervised fine-tuning of LLMs for BDI tasks across linguistically diverse language pairs.
- **Chapter 7: Conclusion and Future Work**: This chapter presents the conclusion of the thesis with few possible future research directions to the thesis.

Chapter 2

Literature Survey

In state-of-the-art cross-lingual word embeddings, researchers have explored various techniques broadly categorized into supervised, weakly supervised, and unsupervised approaches. This primary classification depends on the mode of using a bilingual dictionary or lexicon as supervision. Supervised methods utilize a bilingual dictionary to establish a linear mapping across languages. These methods are further divided into three categories, mapping-based, pseudomixed-based, and joint methods, based on the type of training corpus and the mode of training objective function used to generate shared-space embeddings. Mapping-based methods are further subdivided into Regression, Matrix Factorization, Orthogonal Constraint, Canonical, and Margin-based approaches. This categorization reflects different techniques to compute the transformation matrix W, which aligns semantically similar words across languages in a shared space. Similarly, unsupervised methods are also classified based on how the initial W matrix is formed. The structure of this classification process is illustrated in the Figure 2.1.

2.1 Supervised Method

This method requires a minimum of two monolingual corpora and a bilingual dictionary for a bilingual setting. Monolingual embedding for the two languages is learned through Skip-gram or other approaches like fastText [30]. A bilingual dictionary acts as an alignment or supervision of a classical machine-learning problem.

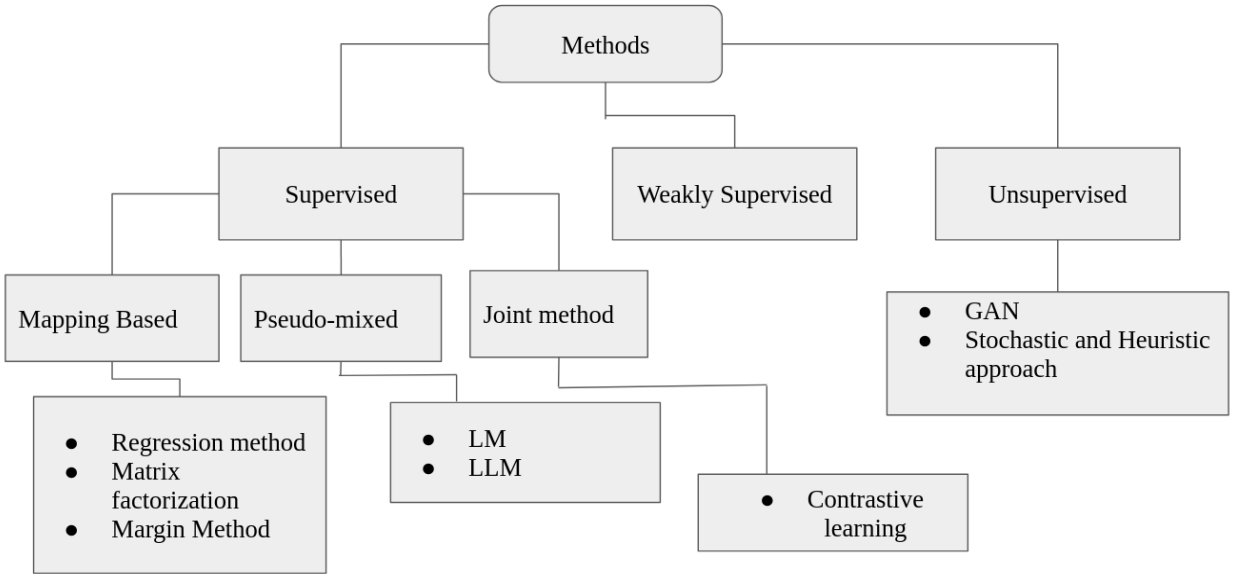


Figure 2.1: Classification of the state-of-the-art method

Figure 2.2: Learning a linear mapping matrix W from the entries of bilingual dictionary X (source language) and Z (target language), then rotate by W to align similar words across languages closer in a shared space

2.1.0.1 Mapping Based

Bilingual dictionary entries and their respective monolingual embeddings are utilized to create a mapping function that aligns semantically similar words across two languages within a shared embedding space. An illustration of this process is shown in Figure 2.2. An important contribution by Thomas Mikolov [8] introduced a regression model to derive this mapping function. The approach originated from the observation that words in different languages exhibit a comparable geometric structure, albeit with varying orientations. An approximate linear mapping is applied to reduce orientation disparities, as depicted in Figure 2.2. The method formulates an objective function as a regression model, which is represented below:

$$\min_{W} \sum_{i=1}^{n} |Wx_i - z_i|^2 \tag{2.1}$$

Here, the objective function aims to minimize the Squared Euclidean distance between the transformed source word embedding Wx_i and the target word embedding z_i derived from the dictionary entries. The optimal transformation matrix W is learned using stochastic gradient descent. For mapping a new source word, x_{new}, its nearest neighbor of Wx_{new} in the target space is identified as the solution. The approach was tested on English-Spanish and Spanish-English bilingual dictionary induction tasks, achieving accuracies of 51% and 52% at P@5, respectively. However, the method faces a significant challenge known as the hubness problem in the nearest neighbor task, where certain elements (hubs) in the target language space become the nearest neighbors for multiple elements in the source language. This phenomenon negatively impacts the performance of the BDI task. Another limitation is the degradation of monolingual properties due to the mapping. Dinu [2] proposed a solution to mitigate the hubness problem by re-ranking the nearest neighbors of test items and reducing the significance of elements with high hub scores. This approach considers the distribution of potential neighbors across numerous mapped vectors. A significant improvement of 12% was observed in English-Italian BDI when trained with 20K dictionary entries. A noteworthy observation is that regularization leads to a decline in BDI accuracy. This reduction is attributed to increased hubness with the application of regularization.

Another approach [31] introduced by Lazaridou also addresses the hubness problem by using a max-margin-based ranking loss in place of squared euclidean distance. The loss for a given pair of training items (x_i,z_i) and the corresponding mapping-based prediction $\hat{z}_i$=Wx_i is given below.

$$\sum_{j \neq i}^{k} max\left\{0, \gamma + dist(\hat{z}_i, z_i) - dist(\hat{z}_i, z_j)\right\} \tag{2.2}$$

Max-margin loss aims to rank the correct translation higher than any other incorrect translation. An interesting observation reported in this work is the significant decrease in the hubness score when compared to the previous approach of Dinu [2] that uses ridge-regression, resulting in a significant increase of 11% accuracy in the BDI task of English-Italian. Both the techniques proposed by Dinu and Lazaridou get the optimal value of W using stochastic gradient descent, and the time complexity is not linear with the vocabulary size of the bilingual training dictionary.

Unlike the previous method, Faruqui and Dyer [20] applied Canonical Correlation Analysis (CCA) to learn a transformation matrix for both source and target languages V and W respectively, thereby bringing the source and the target language in a shared embedding space. CCA maximizes the correlation between the projected vectors Wx_i and Vz_i. Here, a concise cross-lingual representation is learned without increasing the dimension and avoiding irrelevant information that is not generalized across languages. Because of the nature of the transformation, monolingual performance is degraded.

Chao Xing [26] interestingly pointed out inconsistencies among the objective functions of monolingual embedding, cross-lingual learning, and distance or similarity measurement. In the Skip-gram model of monolingual word embedding, the distance measure in training is inner product $c_w^T c_{w'}$, but similarities among word vectors are measured in cosine $\frac{c_w^T c_{w'}}{\|c_w\|\|c_{w'}\|}$. The likelihood function of the Skip-gram model is given below.

$$P(w_{i+j} \mid w_i) = \frac{\exp(c_{w_{i+j}}^T c_{w_i})}{\sum_w \exp(c_w^T c_{w_i})} \tag{2.3}$$

To solve the inconsistency mentioned above, word vectors are normalized to be of unit length, enforcing that the vectors be located in hyperspace. In this case, the inner product is the same as cosine similarity. Similarly, euclidean distance is used in the objective function of linear transformation, but the closeness of words in the projection space is measured in cosine similarity. Here, inconsistency is removed by using cosine distance in linear transformation, $\underset{W}{max} \sum_i (Wx_i)^T z_i$. Wx_i can only be normalized when W is an orthogonal matrix i.e. $WW^T = I$. There is a significant increase of 10% in the BDI task of English to Spanish at P@5 compared to the unnormalized approach. In the word similarity task, the normalized approach shows a higher correlation with the Human rating than the unnormalized word vector. Again, the time complexity in solving W is not linear with the vocabulary size in the bilingual dictionary.

Under the orthogonal constraint, Mikel Artetxe [3] proposed an exact solution $W = VU^T$ that can be computed in linear time with the vocabulary size using matrix factorization technique, Singular Value Decomposition[1] (SVD). Here $U \sum V^T$ is the SVD of $Z^T X$ where X and Z are the matrices form by word embedding of source and translated word, respectively. The details of the closed-form solution are given in Appendix A. The previous method by Xing uses an orthogonality constraint to preserve length normalization. However, this method is motivated by the fact that orthogonal constraint preserves monolingual invariance (preserves length and angle of word vectors). The experimental result also shows that orthogonality is more relevant than length normalization, with an increase of 3% in English-Italian's bilingual dictionary induction task. Performance in monolingual tasks remains the same without degradation as in the original embedding. Previous works show degradation in the performance of monolingual tasks due to restrictions made by transformation. In the later section of this report, we will see a clearer theoretical motivation for enforcing orthogonal constraints in transformation.

[1] https://www.cs.cmu.edu/~venkatg/teaching/CStheory-infoage/book-chapter-4.pdf

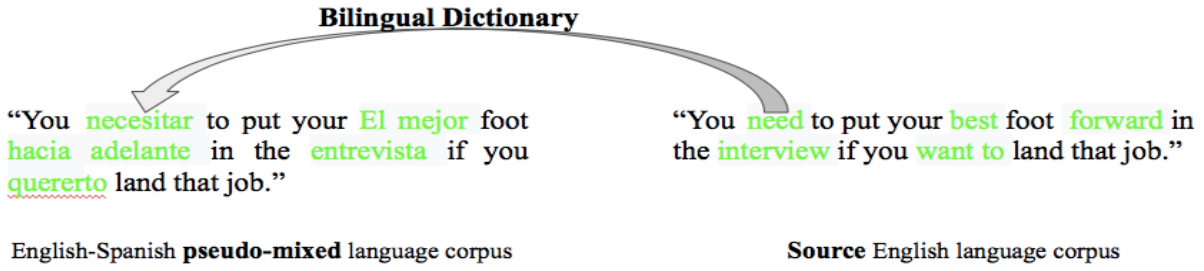


Figure 2.3: A simple illustration of pseudo-mixed corpus

2.2 Pseudo-Mixed Approach

Instead of training multiple (two in a bilingual setting) monolingual word embedding and learning a mapping function between source and target language, other approaches construct a pseudo-bilingual mixed corpus by randomly replacing source word with its translation using word-level alignment from a bilingual dictionary. Then, a bilingual embedding is learned by training on the pseudo-mixed corpus. A simple illustration is presented in Figure 2.3 below.

Min Xiao and Yuhong Guo [32] proposed creating a pseudo-mixed corpus by randomly replacing a source word with its translation word using a seed dictionary. The pseudo-mixed corpus is then trained using a deep neural network such that translation pairs have the same vector representation. The model is evaluated in cross-lingual dependency parsing in eight languages with an average increase of 3.51%.

Gouws and Søgaard [21] explicitly create a pseudo-mixed corpus by concatenating the source and target language corpus followed by a shuffling process. Each word that is part of a translation pair is replaced with its translation equivalent with a probability of $\frac{1}{2k_t}$, where k_t is the total number of possible translation equivalents for a word. A bilingual dictionary is created based on a specific task. In the part-of-speech (POS) class, words like *car* in English and *Maison* in French will be linked as both are nouns. In translation class, the top 20k most frequent words in English are translated into German using Google Translate. Bilingual word embedding is created by training CBOW on this pseudo-mixed bilingual corpus. The model is evaluated on cross-lingual POS tagging in seven languages with an average increase of 3% as compared to the baseline model. One major limitation of the above two approaches is randomly replacing the source word with its target word and training the model without context information.

2.2.1 Joint Approach

In the previous mapping approach monolingual objective is optimized first, followed by optimization of the cross-lingual objective. Joint method Optimizes monolingual and cross-lingual objectives in parallel. Long Duong [22] proposed a method where instead of randomly replacing every word in the corpus with its translation, they replace each center word with a translation. A Pseudo bilingual corpus is created by replacing the middle word (target word) with its translation surrounded by context word. In addition, a method inspired by expectation maximization to handle polysemy explicitly is also proposed. The method chooses as replacement the translation z_i whose representation is most similar to the sum of the source word representation x_i and the sum of the context embeddings x_c. The translation is done using PanLex dictionary[2]. A joint approach is used to predict both the words and their appropriate translations by modifying the objective function with negative sampling in CBOW [4]. The objective function of joint training is given below. In Equation 2.4, $\alpha \log \sigma(u_{w_i}^T h_i)$ is the monolingual objective component, $(1-\alpha) \log \sigma(u_{\bar{w}_i}^T h_i)$ is the cross-lingual objective component and finally $\sum_{j=1}^{p} \mathbb{E}_{w_j \sim P_n(w)} \log \sigma(-u_{w_j}^T h_i)$ is the negative sampling component.

$$J = \sum_{i \in D_e \cup D_f} (\alpha \log \sigma(u_{w_i}^T h_i) + (1-\alpha) \log \sigma(u_{\bar{w}_i}^T h_i) + \sum_{j=1}^{p} \mathbb{E}_{w_j \sim P_n(w)} \log \sigma(-u_{w_j}^T h_i)) \quad (2.4)$$

Method on how to combine the learned embedding V and U is also mentioned. Experimental result on bilingual dictionary induction task shows an average increase of 23% in recall@5 than baseline approach. Evaluation on the monolingual word similarity task also shows improvement over other baseline models. A summary of supervised approach is shown in Table 2.1.

2.2.1.1 Contrastive Learning Approach

To further boost the BDI performance, contrastive learning cross-lingual embedding approach [1] has emerged as a powerful approach that combines static word embeddings (VecMap) with contextual embeddings generated by language models (LMs) and large language models (LLMs). Using a contrastive framework and leveraging dictionary pairs for alignment, this method enhances performance by effectively integrating the strengths of both static and contextual embeddings. The contrastive learning approach offers a promising pathway for improving BDI across a broader range of language pairs. However, the

[2] https://panlex.org

Table 2.1: Summary of supervised approach

Paper	**Language**	**Method**
Mikolov et al.[8]	English, Spanish and Czech	Mapping (Regression)
Dinu et al.[2]	English and Italian	Mapping (Ridge-Regression)
Lazaridou et al.[31]	English and Italian	Mapping (Margin based)
Faruqui and Dyer et al.[20]	English, German, Spanish and France	Mapping (CCA)
Xing et al.[26]	English and Spanish	Orthogonal Constraint
Artetxe et al.[3]	English and Italian	Orthogonal Constraint
Xiao and Guo et al.[32]	Danish Dutch, German, Greek, and Italian,	Pseudo-mixed based
Gouws and Søgaard et al.[21]	Spanish, German, Danish, and Swedish	Pseudo-mixed based
Duong et al.[22]	English, Spanish, Italian and Dutch	Joint method

contrastive learning approach gives lesser performance for linguistic distant and morphologically complex language pairs like English-Manipuri (En-Mn)[33, 34], English-Finnish (En-Fi)[1], and English-Turkish (En-Tr)[1] than similar language pairs like English-Italian (En-It).

2.3 Weakly Supervised

Mikel Artetxe [23] proposed a simple self-learning framework that reduced the need for bilingual dictionaries into as small as 25 dictionary entries. This method exploits the structural similarity of the embedding spaces across languages. With a small seed dictionary, an initial transformation matrix W is learned. The W matrix is used to generate new dictionary entries. Assuming the newly generated dictionary to be of better quality than the initial seed dictionary, it is again used to get a new transformation matrix W, and the process continues till it converges. This method is popularly known as vecmap[3]. A simple illustration of the above method is given below in Figure 2.4. The objective function of this approach is given below.

$$W^* = \arg\min_W \sum_i \sum_j D_{ij} \|X_{i*}W - Z_{j*}\|^2 \tag{2.5}$$

The optimization function in Equation 2.5 can be reformulated as given below, where D is a dictionary matrix such that D_{ij}=1 if x_i is the translation of z_i.

$$W^* = \arg\min_W \mathrm{Tr}(XWZ^TD^T) \tag{2.6}$$

Under orthogonality constraint i.e., $WW^T = I$, the optimal solution of W is $W^* = UV^T$, where $U\sum V^T = X^TDZ$ is the singular value decomposition (SVD) of X^TDZ. The details of the closed form solution is given in Appendix A. The experimental result shows comparable

[3] https://github.com/artetxem/vecmap

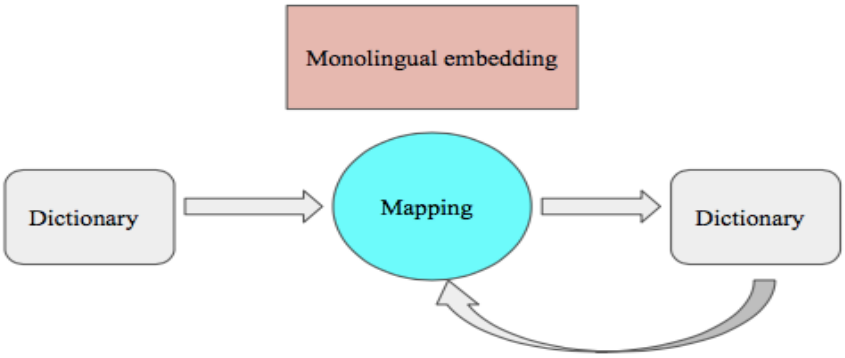


Figure 2.4: A simple illustration of vecmap method

performance with other baseline models that use 5000 dictionary entries in English to Italian bilingual dictionary induction tasks. It also shows that using numerals performance achieve in bilingual dictionary induction task is also comparable with baseline models that use the same 5000 dictionary entries. One major limitation of this method is the assumption of structural similarity of the languages in embedding spaces which might not always be true for distant languages or morphological rich languages.

Samuel L. Smith [35] proposed a new approach without expert bilingual signal by constructing a "pseudo-dictionary" from the identical character strings which appear in both languages. European languages share a large number of words composed of identical character strings e.g. words like "London," "DNA," and "Tortilla" have the same semantic meaning across some European languages. The paper explained that for making the linear transformation to be consistent, orthogonality constraint is imposed. Under consistent linear mapping, If x_i is translated to z_i using W, then z_i should be able to translate back to x_i only when $WW^T = I$. It also proposed an inverted softmax for identifying translation pairs to mitigate the hubness problem. The experimental result shows a significant increase of 10% in precision@5 over baseline model in bilingual dictionary induction task from English to Italian. This method may not be applicable for distant languages that don't share any identical word.

Yerai Doval [24] proposed a method for Aligning bilingual signals in a midway transformation that aims to obtain a better cross-lingual integration of the vector spaces. The idea is to bring closer source words and their translations. This method hypothesizes that the impact of language-specific phenomena and corpus-specific biases will be reduced. This is achieved by least squared optimization between Wx_i and the average of source and its translation $\overrightarrow{\mu_{s,t}} = \frac{\vec{v_s}+\vec{v_t}}{2}$ so that the objective function changed to $\min_W \sum_{(s,t)\in D} \|W\vec{s} - \overrightarrow{\mu_{s,t}}\|_2$. Experimental results show a 2% increase in bilingual induction tasks of English-Spanish as compared to Vecmap [23]. There is also a 7% increase in bilingual dictionary induction of

Table 2.2: Summary of Weak supervised approach

Paper	Language	Method
Artetxe et al.[23]	German, Finnish, English and Italian	Matrix factorization under orthogonal
Smith et al.[35]	English and Italian	Matrix factorization under orthogonality
Yerai et al.[24]	German, Finnish, Spanish, English and Italian	Mapping (Regression)

distant language, English-Finish, as compared to MUSE [36], which is also explained in the later section. A summary of weak supervised approach is shown in Table 2.2

2.4 Unsupervised Approach

With results showing comparable performance in the weakly supervised method as compared to the supervised approach, researchers started to explore in a fully unsupervised way to cross-lingual word embedding. It works on creating an initial transformation matrix W without any bilingual dictionary.

An unsupervised approach [37] introduced by Meng Zhang used an adversarial game [38] to generate the initial transformation matrix W without any bilingual dictionary. It is a two-player game, a discriminator D works to discriminate the two embedding spaces apart, while a generator G works to fool the discriminator by mapping the source language space onto the target language space. The generator G is trained to fool the discriminator D. This is done by letting the generator peak at the losses the discriminator suffers. The generator in our context will be a linear transformation W. The generator chooses W such that its output Wx_i has a distribution close to z_i. The discriminator function to discriminate between vectors z_i and Wx_i. The adversarial model is implemented in three different variations based on the way orthogonality constraint is implemented. The three different models, along with the objective functions, are shown in the Figure 2.5 given below. Model 1 implement orthogonal parameterization [39] of the generator. The orthogonal parameterization is still quite slow. In Model 2, it is relaxed to a self-consistent transformation as in [35]. Another way to relax orthogonal transformation is implemented in Model 3 that implements the constraint as a regularization task.

The experimental result shows a significant increase of 11% with no bilingual seed dictionary over baseline supervised approach with 100 seed dictionary in bilingual dictionary induction task of Chinese to English. Comparable performance with baseline supervised method is achieved in distance languages like English-Finnish (morphological rich language), but the performance is still low. Similar to the previous adversarial game method with a slight modification Alexis Conneau [36] proposed a method to generate an initial alignment matrix

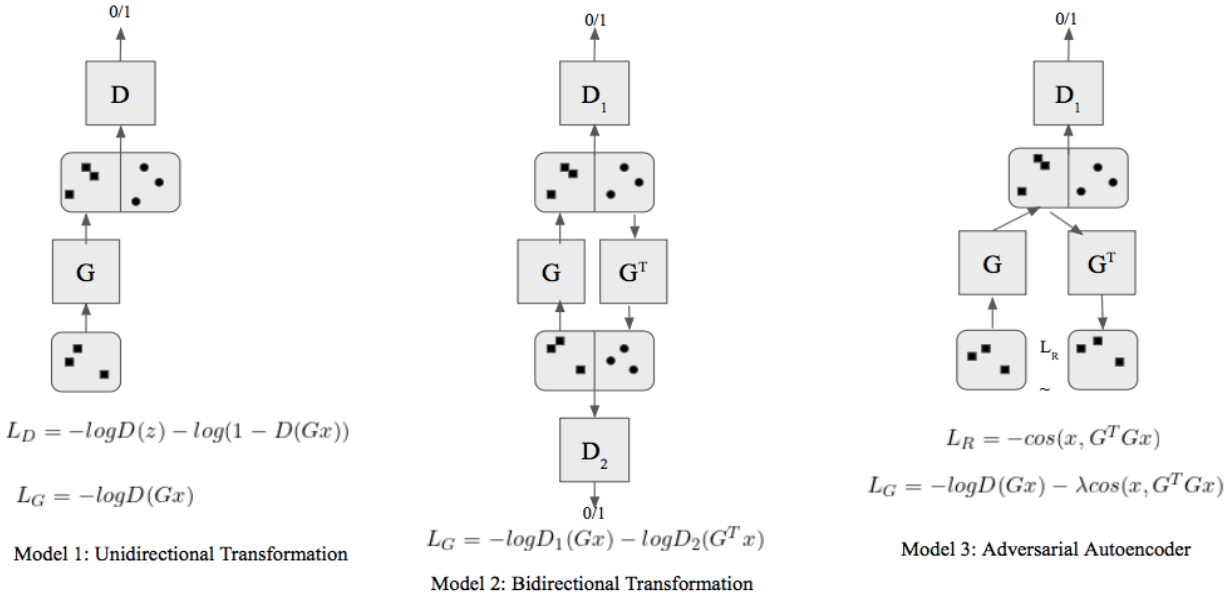


Figure 2.5: Three variation of adversarial models to generate initial transformation matrix W

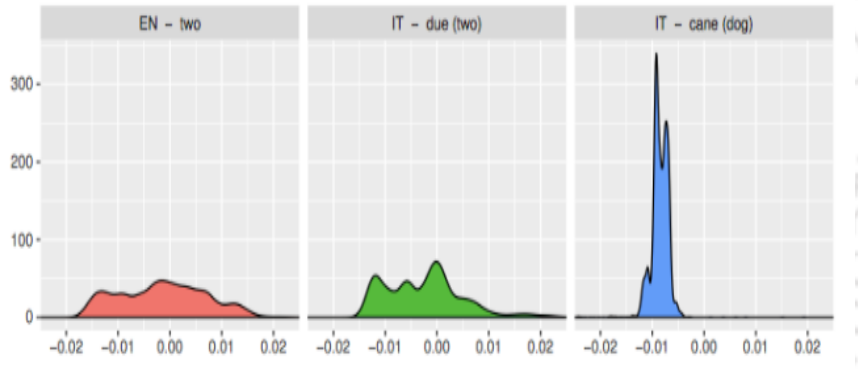


Figure 2.6: Equivalent translations (two and due) have more similar distributions than non-related words (two and cane - meaning dog)

W between embedding spaces using Generative Adversarial network in a domain-adversarial setting [40]. This method is also popularly known as MUSE (Multilingual Unsupervised and Supervised Embedding). The initial alignment W is used to generate D_{syn} synthetic, bilingual dictionary. A generator will randomly sample from $WX = Wx_1, Wx_2, ..., Wx_n$. A discriminator is trained to differentiate the randomly sampled Wx_i from true translation z_i. Mapping objective and discriminator objective are given in equation 2.7 and 2.8 respectively and $P_{\theta_D}(source = 1|z)$ is the probability that a vector z is the mapping of source embedding.

$$L_W(W|\theta_D) = -\frac{1}{n}\sum_{i=1}^{n} \log P_{\theta_D}(source = 0|Wx_i) - \frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(source = 1|z_i) \qquad (2.7)$$

$$L_D(\theta_D|W) = -\frac{1}{n}\sum_{i=1}^{n}\log P_{\theta_D}(source = 1|Wx_i) - \frac{1}{m}\sum_{i=1}^{m}\log P_{\theta_D}(source = 0|z_i) \quad (2.8)$$

D_{syn} is used as a bilingual dictionary to get a better W by refining the bilingual dictionary iteratively in the same way proposed by Artetxe [23]. However, given that the synthetic dictionary obtained using adversarial training is already strong, we only observe small improvements when doing more than one iteration, i.e., the improvements on the word translation task are usually below 1%. Translation with cross-domain similarity local scaling in a bi-partite neighborhood graph is used to mitigate the hubness problem. Experimental result shows comparable performance with supervised baseline model in bilingual dictionary induction task of English to Italian. The result also shows a significant increase of 18% in the same task when comparable corpus Wikipedia is used. Both methods based on the adversarial game often fail on realistic scenarios involving non-comparable corpora and/or distant languages. While adversarial game settings have achieved impressive results at times, they are also highly unstable, e.g., with different initialization leading to precision scores that vary between 0% and 45% for English–Greek [41].

Mikel Artetxe [42] proposed a robust self-learning method based on a fully unsupervised way to get initial bilingual alignment, and a robust self-learning algorithm iteratively improves this solution. The intuition behind the initial bilingual dictionary is two equivalent words in different languages should have a similar distribution obtained the from similarity matrix of all words in the vocabulary. A simple illustration of this idea is given above in Figure 2.6. After generating the initial bilingual dictionary seed lexicon, the vecmap [23] method is used to iteratively improve the seed bilingual dictionary lexicon. Again under orthogonality constraint i.e., $WW^T = I$, the optimal solution of W is $W^* = UV^T$, where $U\sum V^T = X^TDZ$ is the singular value decomposition (SVD) of X^TDZ. D is a dictionary matrix such that D_{ij}=1 if x_i is the translation of z_i. One problem with initial unsupervised dictionary induction is that the quality of this initial dictionary is not good enough to avoid poor local optima. To address the problem, several methods like Stochastic dictionary induction to encourage wider exploration of the search space, restricting the dictionary induction process to the k most frequent words in each language (k=20000), bidirectional dictionary induction, and Cross-domain similarity local scaling (CSLS) to mitigate hubness problem are proposed. The experimental results show comparable performance with other unsupervised methods.The result also shows a significant increase of 18% and 32% in the bilingual dictionary induction task of distant language pair English-Turkish and English-Finnish, respectively. The method is significantly faster than the baseline. The run-time adapts to the difficulty of the task due to the dynamic convergence criterion of the stochastic approach.

Table 2.3: Comparison of various supervised and unsupervised methods on the data-set used in Dinu [2]

Supervision	Method	EN-IT	EN-DE	EN-FI	EN-ES
5k dict	**Mikolov et al. (2013)[8]**	34.93	35.00	25.91	27.73
	Faruqui and Dyer (2014)[20]	38.40	37.13	27.60	26.80
	Dinu et al. (2015)[2]	37.7	38.93	29.14	30.40
	Lazaridou et al. (2015)[31]	40.2	-	-	-
	Xing et al. (2015)[26]	36.87	41.27	28.23	31.20
	Artetxe et al. (2016)[3]	39.27	41.87	30.62	31.40
	Artetxe et al. (2017)[23]	39.67	40.87	28.72	-
	Smith et al. (2017)[35]	43.1	43.33	29.42	35.13
25 dict	**Artetxe et al. (2017)[23]**	37.27	39.60	28.16	-
Init. heurist.	**Smith et al. (2017), cognates[35]**	39.9	-	-	-
	Artetxe et al. (2017),number[23]	39.40	40.27	26.47	-
Unsupervised	**Zhang et al. (2017)[37]**	**0.00**	**0.00**	**0.00**	**0.00**
	Conneau et al. (2018)[36]	45.15	46.83	**0.38**	35.38
	Artetxe et al. (2018)[42]	**48.13**	**48.19**	**32.63**	**37.33**

Table 2.4: Summary of Unsupervised approach

Paper	Language	Method
Zhang et al.[37]	English, Spanish, Italian, Chinese, Japanese and Turkish	Neural Network Model
Conneau et al.[36]	English, Russian, Chinese, Esperanto and Italian	Neural Network Model
Artetxe et al.[42]	German, Finnish, Spanish, English, Italian and Turkish	Stochastic and Heuristic
Megdalena et al.[43]	German, Finnish, Spanish and English	Stochastic and Heuristic

A summary of the evaluation of various supervised and unsupervised methods in the bilingual dictionary induction task in English, Italian, German, Spanish, and Finnish is provided in Table 2.3. The data set used in this evaluation is the same as that used by Dinu [2] (Wacky + Wikipedia + BNC) for the monolingual corpus and Europarl for the bilingual dictionary in the case of the supervised method. The summary of the unsupervised method is shown in Table 2.4.

2.5 Few-shot prompting method in BDI

Recent research has explored the potential of large language models (LLMs) to induce BDI using a few-shot prompting approach[18]. This technique involves leveraging the inherent contextual understanding of LLMs to align embeddings with minimal supervision[44], often achieving better results than traditional VecMap and contrastive learning methods. However, this study has focused mainly on a limited set of resource-rich and structurally similar (closer) language pairs, often neglecting the evaluation of linguistically challenging

and distant language pairs such as English-Manipuri (En-Mn), English-Finnish (En-Fi), English-Turkish (En-Tr), English-Hindi (En-Hi), English-Japanese (En-Ja) and English-Tamil (En-Ta). In addition, evaluation in state-of-the-art LLM models, such as LlaMA, remains largely unexplored, particularly in the aforementioned challenging settings [18].

2.6 CLWE in distant and morphological rich language pairs

An empirical study by Anders Søgaard [41] reports that unsupervised approaches [37] [36] based on adversarial bilingual alignment technique perform very poorly in a morphologically rich language like Finnish in the task of bilingual dictionary induction of English-Finnish. The study also shows how two languages are distant from each other using eigenvector similarity of the nearest neighbor graph of the certain frequent words or nouns of the two languages. Interestingly, the study pointed out a near-perfect correlation between unsupervised bilingual dictionary induction performance and language similarity matrix. Weak supervision from identical words or few bilingual alignments shows more robust performance. Another empirical study by Ivan Vulic [45] shows that even the most robust unsupervised approaches fail in more challenging and distant languages. A series of bilingual lexicon induction (BLI) experiments with 15 diverse languages (210 language pairs) show that fully unsupervised methods still fail for a large number of language pairs. The performance in BLI never surpasses weak supervision with 500-1000 bilingual alignment.

Mozhi Zhang [46] pointed out that orthogonal mapping only works on language pairs whose embeddings are naturally isomorphic. The experimental result shows poor performance in English-Japanese bilingual lexicon induction task using the supervised method. For non-isomorphic pairs, Zhang proposed an Iterative Normalization method to make orthogonal alignment easier by simultaneously enforcing the individual word vectors to be unit of length and zero mean centering. Iterative Normalization consistently improves word translation accuracy in English-Japanese from 2% to 44%.

The above method proposed by Zhang [46] was also extended to the unsupervised approach later. Magdalena Biesialska [43] proposed a self-supervised method to refine the alignment (Dictionary pairs), which was initially generated in an unsupervised way. An initial seed dictionary that was learned in an unsupervised way employs a heuristic initialization method proposed by [42].Then a self-supervised refinement of the alignment is applied after the initial mapping is done. In general, the proposed method is motivated by the assumption that vector spaces of source and target language embeddings have different structures and should

Table 2.5: Summary of dataset used in the state-of-the-art method

S.No	Paper	Corpus name/availability	Bilingual dictionary/availability
1	Mikolov et al.[8]	WMT11/yes	yes
2	Dinu et al.[2]	Wacky+Wikipedia+BNC/yes	Europarl/ yes
3	Lazaridou et al.[31]	Wacky+Wikipedia+BNC/yes	Europarl/ yes
4	Faruqui and Dyer et al.[20]	yes	no
5	Xing et al.[26]	yes	yes
6	Artetexe et al.[3]	Wacky+Wikipedia+BNC/yes	Europarl/ yes
7	Xio and Guo et al.[32]	No	Wikitionary
8	Gouws and Sogaard et al.[21]	yes	yes
9	Duong et al.[22]	yes	yes
10	Artetxe et al.[23]	Wacky+Wikipedia+BNC/yes	Europarl/ yes
11	Smith et al.[35]	Wacky+Wikipedia+BNC/yes	Europarl/ yes
12	Yerai et al.[24]	yes	Europarl/ yes
13	Zhang et al.[37]	yes	no
14	Conneau et al.[36]	yes	Europarl/ yes
15	Artetxe et al.[42]	yes	no
16	Magdalena et al.[43]	yes	Europarl/ yes

not be considered entirely isomorphic. The refinement is done in two phases. The first phase is averaging the vectors, same as propose by Yerai Doval [24].The intuition behind this step is to bring closer source words and their translations. The second phase is Length normalization and means centering. As the entire projection-based unsupervised CLE method relies on the orthogonal assumption, so,to concur with [46] that word embeddings should be of the same unit length. Moreover, they stress the importance of source and target language vector spaces having equal magnitude centers. Experiment on bilingual dictionary induction task of English-Spanish shows comparable performance with supervised baseline method. Results also show comparable performance on the bilingual dictionary induction task of English-Finnish, which are distant languages with different isomorphic structures. A study [47] presents 40 complete morphologically rich dictionaries that span 10 languages, which were manually compiled, and evaluates three state-of-the-art models on their ability to translate less frequent morphological forms. Our results reveal a substantial decline in model performance when handling rare inflections. However, we show that incorporating a simple morphological constraint during training significantly improves results, demonstrating that bilingual lexicon induction systems can benefit from more effective encoding of morphology.

2.7 Comparable corpus in CLWE

Comparable corpora are not an exact translation, but they are two monolingual corpora in different languages that report about the same or similar event.Wikipedia[4] data is an

[4] https://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

example of a comparable corpus. A comparable corpus can be temporal aligned, domain aligned, or event aligned. Gouws and Søgaard [21] use the comparable concept in generating bilingual seed lexicon by aligning words with similar Part-of-speech tags. The comparable concept is later extended in the corpus, where Conneau [36] used it to generate cross-lingual word embedding in an unsupervised way. The result shows a significant increase of 18% in the bilingual dictionary induction task of English-Italian when comparable corpus Wikipedia is used. The additional experiment also shows an increase of 14% on the same task when the baseline supervised method is used along with Wikipedia comparable corpus. One reason behind this gain is due to the similar co-occurrence statistics of Wikipedia corpora. The summary of dataset used in SOTA CLWE is shown in Table 2.5.

2.8 CLWE in Indian languages

Khatri and Bhattacharya [48] perform cross-lingual word embedding in the following languages: Bengali, Telugu, Tamil, Nepali, Sanskrit, Marathi, Punjabi, Malayalam, Gujarati, Konkani, Oriya, Kannada, and Assamese. The method uses Hindi as a source language. The accuracy in bilingual dictionary induction is very less when baseline methods [23] [42] are applied to the above Indian languages. A method called Indic embedding is proposed, which involves converting all languages into a common script, specifically the Devanagari script. The accuracy in the induction of bilingual Hindi-Tamil and Hindi-Malayalam dictionaries is still low.

2.9 Summary

State-of-the-art cross-lingual word embeddings are broadly classified into supervised, weakly supervised, and unsupervised methods, which differ in their reliance on bilingual dictionaries. Supervised approaches learn a linear mapping between monolingual embeddings, using mapping-based, pseudo-mixed, or joint strategies, while addressing challenges such as hubness and potential degradation of monolingual properties. Weakly supervised methods minimize bilingual supervision by iteratively refining a small seed lexicon, while unsupervised techniques employ adversarial training to align embedding spaces without any bilingual signal; however, they often struggle with distant or morphologically rich languages. Furthermore, leveraging comparable corpora and strategies, such as common script conversion in Indian languages, has shown promise; however, balancing performance across distant language pairs remains a challenge.

Chapter 3

Empirical Evaluation on English-Manipuri CLWE

This chapter explores cross-lingual embedding techniques between English and Manipuri, a Tibeto-Burman language characterized as low-resource. Initially, the focus lies on assessing cross-lingual embedding quality through evalua- tion of bilingual dictionary induction across various supervised, weakly supervised, and unsupervised methods. Results indicate suboptimal performance across all state-of-the-art techniques for English-Manipuri cross-lingual embedding. Subse- quently, attention shifts to investigating the influence of linguistic factors, such as Manipuri's morphologically rich nature and differing sentence structure, on Bilingual Dictionary Induction (BDI) task performance. Additionally, the study demonstrates that the sparsity of Manipuri data, relative to English, impacts the quality of cross-lingual embeddings.

3.1 Introduction

The majority of cross-lingual embeddings reported in the literature [8, 20, 21, 3, 22, 23, 24, 43] primarily focus on European languages such as Spanish, Italian, French, and German. These languages exhibit greater similarity to one another [13, 11, 14] compared to non-European languages. A commonly used method, such as mapping, does not perform well for distant language pairs. This limitation may stem from the non-isomorphic nature and significant linguistic differences between the source and target languages.

To address these challenges, some studies have introduced modifications to the existing approaches using stochastic and heuristic techniques [42] for English-Finnish language pairs

and iterative normalization [46] for English-Japanese, English-Chinese, and English-Turkish. Similarly, Khatri and Bhattacharya [48] conducted cross-lingual word embedding studies for several Indian languages, including Bengali, Telugu, Tamil, Nepali, Sanskrit, Marathi, Punjabi, Malayalam, Gujarati, Konkani, Oriya, Kannada, and Assamese, using Hindi as the source language. Their findings indicate that the accuracy of bilingual dictionary induction (BDI) is significantly lower when baseline methods [23, 42] are applied to these Indian languages. Their proposed method introduces an Indic embedding by converting all languages into a common script, the Devanagari script. However, the BDI accuracy for Hindi-Tamil and Hindi-Malayalam remains relatively low. Furthermore, other methods, such as iterative normalization [46], used for English-Hindi, also demonstrate comparatively low performance in BDI.

In this chapter, we conduct an empirical investigation into the performance of various state-of-the-art cross-lingual embedding methods for Manipuri-English language pairs, identifying challenges in developing efficient methods for this language pair. Our research explores cross-lingual embeddings in Manipuri (Meiteilon), a low-resource Tibeto-Burman language spoken primarily in Manipur, a northeastern state of India. As a distant language from English [9], Manipuri exhibits significant linguistic differences, including a complex morphological structure [16, 49, 17] and a distinct sentence structure [16, 49, 17].

This research serves as an initial empirical study on cross-lingual word embedding between English and Manipuri using various state-of-the-art methods. We examine both supervised and unsupervised approaches to assess the feasibility of minimal or no supervision. Given the scarcity of parallel aligned corpora for English and Manipuri, our empirical study explores the feasibility of comparable corpora, which, unlike direct translations, describe the same event in different languages. Additionally, we investigate the efficacy of cross-lingual embedding in structurally distinct languages such as English and Manipuri by considering linguistic features such as morphology and sentence structure. The study also evaluates the significance of selecting dictionary pairs for training and examines the impact of increasing data size on embedding quality.

Experimental results indicate that English-Italian cross-lingual embeddings, where languages are more closely related, perform significantly better than English-Manipuri embeddings. Following standard practices in the literature, we use top-k frequent training dictionary pairs, considering only the source word frequency, for evaluation. Error analysis reveals that the rich morphological characteristics of Manipuri negatively impact cross-lingual embedding quality. Additionally, the frequency imbalance between English source words and Manipuri target words in dictionary pairs further degrades performance. The results also demonstrate that word order in the corpus influences the Bilingual Dictionary Induction (BDI) task.

3.1.1 Contributions

The key contributions of this chapter are as follows:

1. This chapter provides a thorough empirical analysis of cross-lingual word embeddings (CLWE) between English and Manipuri for bilingual dictionary induction (BDI), employing both state-of-the-art supervised and unsupervised methods.
2. It highlights the linguistic complexities of the Manipuri language that contribute to suboptimal CLWE performance.
3. This chapter demonstrates how an imbalanced frequency distribution of source and target words in the English-Manipuri bilingual dictionary adversely affects cross-lingual alignment.
4. It shows that the rich morphological structure of the Manipuri language poses challenges to the effectiveness of CLWE in the BDI task.

3.2 Related work

In many studies on cross-lingual word embeddings, researchers have explored different techniques, generally classified into supervised, weakly supervised, and unsupervised methods. This categorization is primarily based on how a bilingual dictionary or lexicon is employed to guide and oversee the embedding process.

3.2.1 Supervised

A supervised approach introduced in [8] employs a regression model to derive the mapping function. This method applies an approximate linear transformation to minimize orientation differences, represented as $\min_W \sum_{i=1}^{n} |\vec{x_i}W - \vec{z_i}|^2$, where x_i and z_i denote the source and target word embeddings, respectively. Dinu [2] later enhanced this by incorporating ridge regression. Another refinement [31] introduced a max-margin-based ranking loss instead of squared Euclidean distance. The max-margin loss function prioritizes ranking the correct translation above any incorrect alternatives. Both Dinu's and Lazaridou's approaches determine the optimal transformation using stochastic gradient descent; however, this results in a time complexity that does not scale linearly with the vocabulary size of the bilingual training dictionary.

In contrast, Faruqui and Dyer [20] utilized Canonical Correlation Analysis (CCA) to learn a transformation matrix for both source and target languages, aligning them in a shared embedding space. CCA operates by maximizing the correlation between the projected vectors. Meanwhile, Chao Xing [26] identified inconsistencies between the objective functions of monolingual embeddings, cross-lingual learning, and similarity measurement. To address this, word vectors were normalized to unit length, ensuring their distribution within a hypersphere. Under an orthogonality constraint, Mikel Artetxe [3] proposed an exact closed-form solution, $W = VU^T$, which can be computed in linear time concerning vocabulary size using matrix factorization via Singular Value Decomposition[1] (SVD). Here, $U \sum V^T$ represents the SVD of $Z^T X$, where X and Z are matrices composed of the source and translated word embeddings.

Another set of techniques constructs a pseudo-bilingual corpus by randomly replacing source words with their translations based on word-level alignment from a bilingual dictionary. This allows training a bilingual embedding model on the pseudo-mixed corpus. Gouws and Søgaard [21] explicitly formed a pseudo-mixed corpus by concatenating the source and target language corpora, followed by a shuffling process. For words in a translation pair, replacement with their translation equivalent occurs with a probability of $\frac{1}{2k_t}$, where k_t represents the total number of possible translations for a given word. Long Duong [22] proposed a variation in which, instead of replacing every word randomly, only the center word in a given context is substituted with its translation. A joint learning approach was employed, modifying the CBOW [4] objective function by incorporating negative sampling to predict both words and their translations.

Recent contrastive learning techniques have demonstrated potential in improving cross-lingual word representation [1]. The method proposed in [1] applies contrastive learning specifically to positive word pairs (dictionary-aligned words) and negative word pairs (non-dictionary pairs). However, its effectiveness is constrained when dealing with morphologically complex languages such as Finnish and Turkish.

3.2.2 Weakly Supervised

Mikel Artetxe [23] introduced a weakly supervised self-learning framework, known as vecmap[2], which significantly reduced the dependency on bilingual dictionaries, requiring as few as 25 dictionary entries. Samuel L. Smith [35] proposed an alternative approach that eliminates

[1] https://www.cs.cmu.edu/~venkatg/teaching/CStheory-infoage/book-chapter-4.pdf
[2] https://github.com/artetxem/vecmap

Table 3.1: Statistics of English-Manipuri (En-Mn) comparable corpus and English-Italian (En-It) parallel corpus

Data set	Platform	sentences		words		unique words	
		EN	**MNI**	**EN**	**MNI**	**EN**	**MNI**
En-Mn	Sangai Express	129546	181553	3.5M	3.3M	15247	24449
	Poknafam	55707	86049	1.5M	1.6M	9682	14305
En-It	European Parliament	**EN**	**IT**	**EN**	**IT**	**EN**	**IT**
		1.9 M	1.9M	49M	47M	151017	219976

the need for an expert-provided bilingual signal by generating a "pseudo-dictionary" using identical character strings occurring in both languages. Yerai Doval [24] introduced a method that aligns bilingual signals through a midway transformation, enhancing the cross-lingual integration of vector spaces by bringing source words and their translations closer together.

3.2.3 Unsupervised

An unsupervised approach introduced by Meng Zhang [37] employed an adversarial game [38] to learn the initial transformation matrix W without requiring a bilingual dictionary. Building upon this adversarial game framework with a slight modification, Alexis Conneau [36] proposed a method to obtain the initial alignment matrix W between embedding spaces using a Generative Adversarial Network in a bilingual-adversarial setting [40]. This approach is commonly referred to as MUSE (Multilingual Unsupervised and Supervised Embedding). Meanwhile, Mikel Artetxe [42] introduced a robust self-learning method, which first derives an initial bilingual alignment in a fully unsupervised manner and then iteratively refines the solution through a self-learning algorithm.

3.3 Dataset

Most cross-lingual research in the literature relies on parallel corpora for analysis. However, Manipuri, as a low-resource language, lacks a sufficiently large English-Manipuri parallel corpus. To address this limitation, our empirical study utilizes a comparable corpus, which consists of documents or articles that, while not direct translations, report on the same or similar events in different languages. A well-known example of such a corpus is Wikipedia[3]. We extend the comparable corpus employed in previous studies [9, 15, 50], with specific details outlined in Table 3.1. In particular, the English-Manipuri comparable corpus is

[3] https://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

Table 3.2: Details of the methods for empirical evaluation

Method	**paper**	**method**
supervised	Mikolov 2013[8]	Regression
	Artetxe 2016[3]	Matrix Factorization (VecMap)
	Artetxe 2017[23]	VecMap with refinement
	MBERT [55]	Masked Language Modelling
	Zhang 2019 [46]	iterative normalization
	Li 2022[1]	Contrastive learning refinement
semi-supervised	Artetxe 2017[23]	Orthogonal Mapping with refinement
unsupervised	Zhang et.al 2017[37]	GAN[38]
	Artetxe et.al 2018 [42]	Heuristics method
	Conneau et.al 2017 [36]	GAN[38])+refinement
	Magdalena et.al 2020 [43]	Heuristics method+midway mapping

compiled from two widely recognized Manipuri online news platforms: Sangai Express[4] and Poknafam[5].

Supervised cross-lingual embedding methods generally require a substantial number of dictionary pairs for effective cross-lingual alignment. However, since Manipuri is a low-resource language, it lacks a large set of bilingual dictionary pairs, and the available ones must also be present in the corpus under study. To mitigate this issue, we leverage transliteration pairs of loan words and Named Entities. These translation pairs consist of words that share phonetic similarities and semantic meanings while being written in different scripts. Moreover, transliteration pairs of loan words and Named Entities are expected to exhibit semantically similar neighboring words. These transliteration pairs are extracted from the comparable corpus using a sequence-to-sequence[51] auto-encoder with a grapheme-based representation[52], as detailed in [53]. Additionally, we manually curate transliteration pairs of loan words (or technical terms) and Named Entities.

For the English-Italian language pair, we use the Europarl[6] parallel corpus [54], which consists of translated proceedings from the European Parliament in 21 European languages, including English and Italian. Specifically, for English-Italian, we adopt dictionary pairs from the MUSE[7] library [36], ensuring that the number of dictionary pairs used for English-Manipuri is matched to maintain a fair comparison in our experiments.

3.4 Experimental Setups

To assess the quality of cross-lingual embeddings between English and Manipuri, a preliminary empirical study was conducted using four state-of-the-art supervised methods: Mikolov et al. 2013 [8], Artetxe et al. 2016 [3], Artetxe et al. 2017 [23], and Zhang et al. 2019 [46]. Additionally, a weakly supervised approach, Artetxe et al. 2017 [23], was assessed using 25 English-Manipuri dictionary pairs. Furthermore, four unsupervised methods were evaluated without English-Manipuri dictionary pairs: Zhang et al. 2017 [37], Conneau et al. 2017 [36], Artetxe et al. 2018 [42], and Magdalena et al. 2020 [43]. The weakly supervised approach [23] was also tested in a supervised setting incorporating self-learning. Lastly, the English-Manipuri cross-lingual embedding was evaluated on a Multilingual-BERT (MBERT) architecture [55] with cross-lingual alignment, where source words were substituted with target words using a bilingual dictionary within the combined source and target corpus. The next sentence prediction module in the MBERT architecture was ignored. The aforementioned methods were also evaluated for the closely related European language pair, English-Italian. Details of the methodologies are summarized in Table 3.2.

In this study, monolingual word embeddings for English, Manipuri, and Italian corpora were generated using Word2Vec with the Continuous Bag of Words (CBOW) model [4], employing a dimensionality of 300, a window size of 5, and ignoring words occurring fewer than ten times. All state-of-the-art methods referenced above were evaluated using these embeddings. The training dataset comprised 3900 word pairs, while 650 pairs were reserved for testing. The same experimental setup was replicated for the English-Italian language pair. These methods were assessed using the Bilingual Dictionary Induction (BDI) task with metrics P@1 (Precision at 1) and P@5 (Precision at 5). P@1 considers only the nearest neighbor, while P@5 accounts for up to the fifth closest neighbor based on the ground truth. All listed methods in Table 3.2 learn a transformation matrix W of size 300×300, aligning with the monolingual embedding dimensionality. Methods in [23, 42] employ an iterative self-learning procedure until convergence is achieved.

For the approach outlined in [14], iterative normalization is followed by five iterations of the Procrustes refinement method [36]. Iterative normalization simultaneously ensures normalization and mean-centering. In the adversarial-based unsupervised approach [37], dictionary pairs are generated with a single discriminative hidden layer of dimension 500. The unsupervised method in [42] is evaluated with an unsupervised vocabulary size of 4000, source

[4] https://www.thesangaiexpress.com/index.html
[5] http://www.poknapham.in
[6] https://www.statmt.org/europarl/
[7] https://github.com/facebookresearch/MUSE

reweighting of 0.5, target reweighting of 0.5, self-learning enabled, and a vocabulary cutoff of 20000 words. The approach in [36] employs adversarial training with two discriminator layers of 2048 dimensions, using embeddings from the 75000 most frequent words. The adversarial network undergoes five training epochs, with 1000000 iterations per epoch and a batch size of 32. The refinement mode training involves five refinement iterations. The unsupervised method described in [43] follows the same setup as [42] to generate dictionary pairs in an unsupervised manner.

For evaluating multilingual BERT (MBERT)[55], the EMBERT (English-Manipuri BERT) architecture consists of two hidden layers with a hidden size of 128 and a maximum sequence length of 512. The average embedding is computed over varying contexts. Once EMBERT embeddings are obtained, they are segregated into separate English and Manipuri embedding spaces, and BDI evaluation is conducted using cosine similarity between the two spaces. The contrastive fine-tuning approach[1] is evaluated with the following hyperparameters: $N_{iter} = 5$, $N_{+ve} = 50$, and $\lambda = 0.3$, where N_{iter} denotes the number of iterations in VecMap. Fine-tuning of static word embeddings (WEs) is performed using an SGD optimizer with a learning rate of 1.5, while large language model (LLM) WEs are fine-tuned using the AdamW optimizer [56] with a learning rate of $2e^{-5}$. Both static (VecMap CLWE) and LLM WEs are fine-tuned for five epochs with $\tau = 0.1$. For English and Italian, LLM WEs are extracted from $mBERT_{base}$[55], while Manipuri LLM WEs are obtained from IndicBERTv2-MLM-only[57].

3.5 Result and Analysis

From Table 3.3, it can be seen that En-It outperforms En-Mn by a significant margin across supervised, unsupervised, and weakly-supervised settings.

3.5.1 Supervised Approach

The contrastive fine-tuning approach achieved superior performance over all supervised methods in both P@1 and P@5 for En-Mn. Similarly, for En-It, contrastive learning outperformed all supervised methods at both P@1 and P@5. The method utilizing mBERT embeddings demonstrated lower performance compared to other approaches for both En-Mn and En-It. This could be due to the reliance on contextual word information, which primarily aids in word alignment within the cross-lingual embedding space when accompanied by contextual alignment information in bilingual dictionaries.

Table 3.3: Result of evaluation on Supervised, Weakly-supervised and Unsupervised method on English-Manipuri and English-Italian

Semi-supervised	**En-Mn**		**En-It**	
	P@1	**P@5**	**P@1**	**P@5**
Artetxe 2017 [23] 25 dictionary	08.00	17.85	29.29	49.71
Artetxe 2017 [23]numerals	00.15	00.15	36.14	52.71
Unsupervised	**P@1**	**P@5**	**P@1**	**P@5**
Zhang 2017 [37]	00.00	00.00	30.45	44.53
Conneau 2017 [36]	00.00	00.00	46.45	57.25
Artetxe 2018 [42]	**04.46**	**12.15**	41.14	54.28
Biesialska 2020 [43]	00.23	01.78	42.23	58.76
Conneau 2017 [36]+iterative normalization	00.00	00.00	48.09	58.78
Supervised	**P@1**	**P@5**	**P@1**	**P@5**
Mikolov 2013 [8]	01.08	02.77	18.43	33.00
Artetxe 2016 [3]	12.15	23.08	46.00	65.29
Artetxe 2017 [23]	09.23	19.08	38.57	55.29
Iterative normalization [46]	13.09	23.73	48.43	65.43
mBERT [55]	00.14	00.14	11.00	15.28
Li 2022 [1]	**17.71**	**33.43**	**54.57**	**70.28**

3.5.2 Unsupervised Approach

Unsupervised approaches utilizing GANs (Generative Adversarial Networks) [37], [36], and [36] combined with the Iterative Normalization method fail entirely in the En-Mn setting. This failure is likely attributed to the poor initial alignment generated by the GAN-based methods. In contrast, the unsupervised VECMAP approach [42], which leverages structural similarities, surpasses all other unsupervised techniques, achieving P@1=4.16 and P@5=12.15. For En-It, the mid-way alignment strategy [43] demonstrates superior performance over all other unsupervised methods.

3.5.3 Weakly Supervised Approach

The weakly supervised approach utilizing numerals demonstrates significantly lower performance in En-Mn compared to En-It. This poor outcome may be attributed to the fact that En-It shares the same numeral system, which is not the case for En-Mn.

Table 3.4: Five nearest predicted translation of English word. The English meaning of the predicted words are given in the next row.

English word	Baseline predicted	Ground truth
musa	মুসাগী, মুটা, মুসু, মুতা, মুসানা for musa, muta, musu, muta, by musa	**মুসা**
give	পীথোকপা, পীবদি, পীবসু, লৌরকপা, লৌবা give away, to give, to give, took it from, take	**পীবা**
nothing	তশেংনা, করিনো, করিগী, করিগীনো, করিগীদমক truthfully, what, why, why, for what reason	**করিমত্তা**
colour	পেকেত, মচুগী, সুগর, হেরোইন, খাঙেন packet, colour's, sugar, heroin, gold measurement unit	**মচু**
essential	তঙাইফদনা, চাননবা, শীজিন্ননবা, লৌমীশিংদা, পোৎলমশিং critical, matching, to use, to farmers, goods	**তঙাইফদবা**

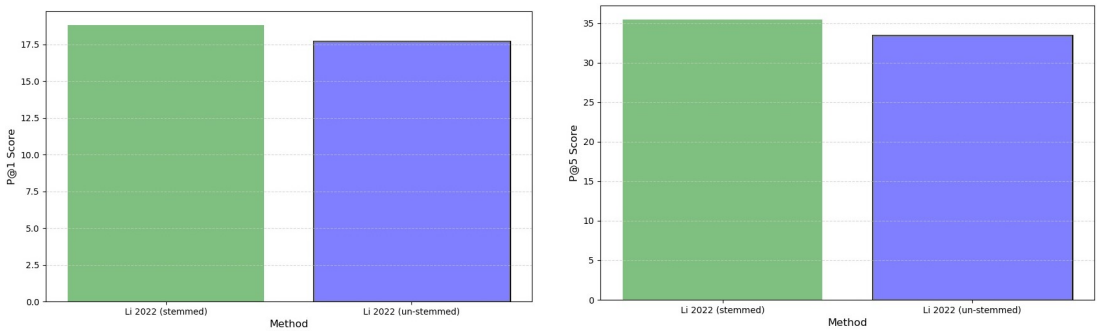
From these observations, it is evident that the quality of cross-lingual embeddings generated through supervised, weakly supervised, and unsupervised methods for distant language pairs like English-Manipuri is considerably inferior to that of closely related languages such as English-Italian when evaluated under the same experimental setup. This substantial performance gap arises due to linguistic differences, including morphological variations between the source and target languages, as well as disparities in word order.

3.5.4 Error Analysis

For error analysis, we consider the results obtained using the contrastive fine-tuning approach proposed by [1], a supervised method. Manipuri, being a morphologically complex language, exhibits meaning variations due to the addition of prefixes or suffixes. The BDI task evaluated in this study is impacted by this morphological richness, where target words with affixes tend to move closer to the source word, thereby pushing the ground truth target word further away.

As shown in Table 3.4, the English word **musa** is translated into Manipuri as **মুসাগী**. However, the presence of the suffix **গী** introduces a slight semantic shift. Similarly, the first nearest translation of **essential** includes the inflection **না**, altering its meaning from **essential** to **critical**. The words **give**, **nothing**, and **colour** are also translated into Manipuri as **পীবসু**, **করিনো**, and **মচুগী**, respectively. Notably, the predicted Manipuri translations incorporate suffixes: **সু** in **পীবসু**, **নো** in **করিনো**, and **গী** in **মচুগী**. The details of the five nearest Manipuri translations of an English word, along with their English meanings and the ground truth Manipuri translations, are provided in Table 3.4.

To validate this observation, we apply a Manipuri suffix segmenter based on the widely used unsupervised GRAph-based Stemmer [58] to stem the Manipuri data. The results reveal



P@1 P@5

Figure 3.1: Comparative analysis of Li 2022[1] methods in stemmed and un-stemmed

Table 3.5: Correlation analysis in training and testing dictionary

Data	Training pairs	Success training pairs	Failure training pairs
English-Italian	0.72	0.75	0.06
English-Manipuri	0.15	0.50	0.07

a notable P@5 improvement of 6.07% when evaluated using [1], reinforcing the conclusion that the rich morphological structure of Manipuri negatively affects BDI task performance.

3.5.5 Imbalance Frequency Distribution of words in (source, target) pairs

In the training of state-of-the-art methods, dictionary pairs are selected based solely on the frequency of the source word, without considering the frequency of the target word. To assess the significance of both frequencies, the correlation between the source and target word frequencies—ranked in descending order of source word frequency—was computed for English-Italian and English-Manipuri. This correlation was analyzed for both training and testing datasets. The correlation for the top-4200 dictionary pairs in English-Manipuri is 0.15, which is significantly lower than that of English-Italian (0.71) at the same threshold. Moreover, correlation analysis on the testing data reveals that successful pairs exhibit a much stronger correlation than failed pairs in both English-Manipuri and English-Italian. The detailed results of this correlation analysis are presented in Table 3.5. These findings strongly indicate that aligning source and target words with comparable frequencies could enhance the quality of cross-lingual embeddings for English-Manipuri, as the higher correlation observed in English-Italian corresponds to improved performance.

Table 3.6: Five nearest predicted translation of English word. The English meaning of the predicted words are given in the next row.

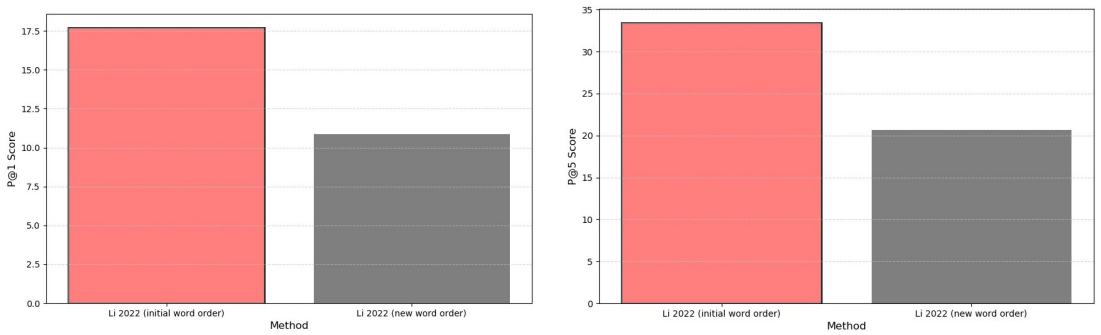
English word	Baseline predicted	Ground truth
stress	ফত্তবা, লায়নাশিং, শিংনবশিং, প্রোব্লেম, খুদোংথিবা feel, diseases, grudges, problem, danger	**স্ত্রেস**
maring	মরিং, রোংমৈ, আইমোল, কবুই, তাঙ্খুল maring, rongmei, aimol, kabui, tangkhul	**মরিং**
philosopher	লম্বোইবা, ফিলোসোফর, কবি, শৈশকপা, স্বামি saint, philosopher, poet, singer, swami	**ফিলোসোফর**
thongkhong	নাগামপাল, বাজারদগী, ককৱা, অরোং, নাওরেমথোং nagamapal, from bazar, kakwa, arong, naoremthong	**থোংখোং**
pena	ফোল্ক, জগোই, আর্টিষ্ট, মাইবী, খুনুং flute, dance, artist, oracle, folk	**পেনা**

3.5.6 Grouping of semantically similar word which are not direct translation

Cross-lingual embedding also group together semantically related words which are not direct translation. A few e.g is given in Table 3.6. These examples are generated by evaluating in contrastive fine-tuning method[1]. Five nearest neighbour of the word **stress** in English translation are **feel**, **diseases**, **grudges**, **problem** and **danger**. The corresponding Manipuri translation can be refer from the same Table 3.6. This suggest that the quality of cross-lingual embedding between English-Manipuri may vary depending on different evaluation process. The CLWE generated may not be good for BDI tasks due to direct comparison of English word and ground truth Manipuri word in the given test pairs. However, it may serve purpose to other extrinsic downstream task like Machine Translation and Sentiment classification.

3.5.7 Difference in word order

English and Manipuri exhibit distinct word order patterns, with English adhering to the **subject + verb + object** structure, while Manipuri follows the **subject + object + verb** order. To investigate the impact of word order on cross-lingual embedding quality, the words in each sentence of the Manipuri dataset were randomly rearranged. The newly reordered data was then evaluated using the contrastive learning method proposed by [1]. As depicted in Figure 3.2, the results indicate a significant decline in performance when word order is altered, demonstrating that word order plays a crucial role in cross-lingual word embedding. Additionally, its quality may also depend on the approach used to generate distinct monolingual embeddings.



P@1 P@5

Figure 3.2: Comparative analysis of Li 2022[1] methods in initial word order and randomly changed word order

3.5.8 Zipf's Plot

Zipf's law states that frequency decreases rapidly with rank and follows an approximate power-law relation. The plot in Figure 3.3 illustrates that the slope for Manipuri data is steeper than that of English data, indicating greater sparsity in the Manipuri dataset. The sparse nature of the Manipuri data implies that the available vocabulary for training cross-lingual embeddings is limited. Consequently, the embeddings may fail to capture the full semantic richness of the language, resulting in incomplete representations of words and concepts. Sparse data lacks sufficient contextual information necessary to capture the nuanced meanings and usage patterns of words. This limitation can lead to embeddings that do not effectively represent the semantic relationships between words in Manipuri and their counterparts in other languages. Additionally, the sparsity of the data may hinder machine learning models from learning meaningful associations between Manipuri words and their translations, leading to inaccurate or incomplete cross-lingual embeddings that fail to reflect the true semantic similarities between languages. Furthermore, a higher degree of sparsity increases the likelihood of noise or irrelevant information within the training data, which can disrupt the learning process and degrade the accuracy of cross-lingual embeddings.

3.6 Summary and Future work

Based on our experimental findings and error analysis, it is evident that the linguistic disparity between English and Manipuri presents significant challenges in achieving high-quality cross-lingual embeddings, particularly when evaluated in bilingual dictionary induction (BDI) tasks. The sparsity of Manipuri data, a consequence of its complex morphological structure, further exacerbates these difficulties. Beyond being a morphologically rich language, Manipuri follows a Subject-Object-Verb (SOV) sentence structure, differing from

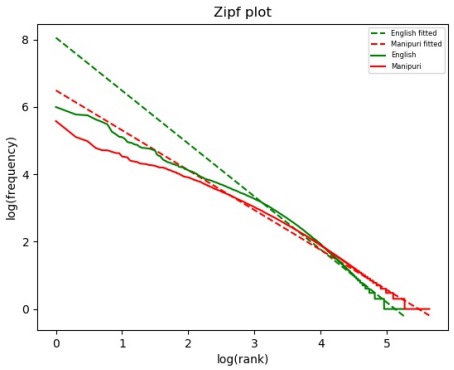


Figure 3.3: English vs Manipuri data Zipf's Law plot

English's Subject-Verb-Object (SVO) order, which negatively impacts the quality of cross-lingual word embeddings (CLWEs).

Our experiments further demonstrate that unsupervised Generative Adversarial approaches fail to perform effectively for linguistically diverse language pairs such as English-Manipuri. This inefficacy is likely attributed to the sub-optimal initial alignment generated by the Generative Adversarial Network (GAN). Additionally, leveraging multilingual BERT embeddings—despite their rich contextual representations—degrades performance in BDI tasks. This may be due to the lack of contextual alignment in the training dictionary pairs.

Moving forward, we intend to assess the quality of the bilingual dictionary pairs used in training. Investigating BDI performance across varying training dictionary sizes may provide valuable insights into cross-lingual alignment effectiveness. Furthermore, expanding the English-Manipuri comparable corpus is planned to enhance training data richness. It is also noteworthy that cross-lingual embeddings tend to group semantically similar words, even when they are not direct translations. Consequently, beyond BDI, we aim to explore alternative evaluation tasks, such as machine translation and sentiment classification, to gain deeper insights into the effectiveness of cross-lingual embeddings.

Experimental observations further indicate that linguistic diversity between English and Manipuri significantly hinders the quality of cross-lingual embeddings in the BDI task. The sparsity of Manipuri data, resulting from its complex morphological structure, plays a crucial role in this challenge. Notably, merely increasing dataset size does not necessarily improve CLWE quality in BDI, especially if the additional data is sparse and has a skewed frequency distribution. In such cases, the embedding quality may either degrade or remain unchanged, depending on the distribution of added data. This indirectly suggests that CLWE performance is domain-dependent, as different domains exhibit distinct frequency distributions and word co-occurrence structures.

A key insight from our empirical analysis highlights that the training dictionary should ensure comparable coverage of both source and target words. This suggests that the semantic neighbors of a source word should be similar to those of the corresponding target word. Instead of frequency-based dictionary selection, an approach leveraging centrality measures to balance source-target word coverage could yield better results. Another crucial finding is that CLWE quality depends not only on the mapping function and dictionary pairs but also on the monolingual embedding generation method. Future work could explore joint learning approaches that better capture cross-lingual contextual information to enhance embedding quality.

Finally, given that cross-lingual embeddings also cluster semantically similar words that are not direct translations, it would be valuable to explore evaluation tasks beyond BDI, such as cross-lingual document classification and sentiment classification, to further examine the robustness and effectiveness of these embeddings.

Chapter 4

Improving Linear Orthogonal Mapping approach

Orthogonal linear mapping is a widely used technique for generating cross-lingual embeddings between two monolingual corpora, relying on word frequency-based seed dictionary alignment. While effective for isomorphic language pairs, its performance degrades when applied to distant language pairs with distinct sentence structures and morphological characteristics. For such distant language pairs, existing frequency-aligned orthogonal mapping methods face two key challenges: (i) discrepancies in word frequency distributions between the source and target languages, and (ii) varying contributions of different word pairs in the seed dictionary. To address these issues, this chapter introduces a novel centrality-aligned ridge regression-based orthogonal mapping approach. The proposed method employs centrality-based alignment for selecting the seed dictionary and incorporates a ridge regression framework to assign influential weights to different word pairs. Experimental evaluations conducted on five language pairs—spanning both isomorphic and distant languages—demonstrate that the proposed method surpasses baseline approaches in Bilingual Dictionary Induction (BDI), Cross-lingual Sentence Retrieval Task (CSRT), and Machine Translation. Additionally, extensive analyses are provided to further validate the effectiveness of the proposed approach.

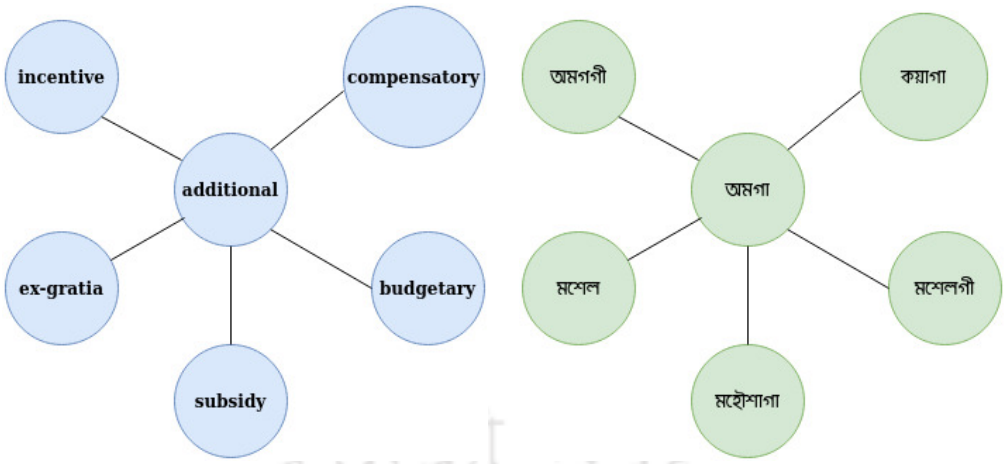


Figure 4.1: English vs Manipuri translation with five Nearest Neighbour

4.1 Introduction

Recently, CLWE has gained considerable attention in research for advancing various cross-lingual NLP applications, including text classification [59], information retrieval [60], machine translation [61], question answering [62], cross-lingual word analogies [63], cross-lingual relation extraction [64], cross-lingual article linking [65], emotion detection [66], and sentiment classification [67]. While there has been recent exploration of contextualized representation methods [68], *projection-based CLWE* methods remain among the most effective, particularly for the bilingual dictionary induction (**BDI**) task, when compared to multilingual contextual representations such as mBERT [69]. Embeddings extracted from multilingual LLMs like mBERT lack the lexical properties of static embeddings such as Word2Vec and FastText. Since LLMs generate embeddings at the token level, their representations are influenced by sentence context and token positioning. Additionally, sub-word tokenization can degrade lexical information [70]. Projection-based CLWE methods, which learn mappings from monolingual embeddings with limited or no supervision [71], typically employ orthogonal transformations based on l_2 norms to align source and target language embeddings.

The initial linear projection approach based on l_2 norms was introduced by [72], later refined by [26] through the introduction of orthogonality constraints to maintain vector length and angle consistency between source and target words. While these approaches utilized gradient descent for learning projection matrices, [3] proposed a closed-form solution via Singular Value Decomposition (SVD) with orthogonal constraints. This method was further improved in [23] by incorporating iterative refinement, reducing dependency on extensive bilingual dictionaries.

Although these methods perform effectively for isomorphic language pairs, as noted in [12], they exhibit limitations when applied to non-isomorphic language pairs. Our empirical analysis highlights two key challenges in handling non-isomorphic language pairs:

- **Selection of seed dictionary pairs:** Traditional orthogonal mapping methods [26, 3, 23, 12] typically select seed bilingual dictionary pairs based on the most frequent words in the source language, assuming comparable word frequencies across source and target corpora. However, this assumption does not hold for non-isomorphic languages (as detailed in Section 4.4.1.2). Our study demonstrates that frequency-based dictionary selection is suboptimal for non-isomorphic language pairs. Instead, we find that word centrality within a co-occurrence network is more comparable across languages than frequency. Network centrality measures the importance of a node based on its topological properties, making it a more reliable criterion for selecting seed dictionary pairs.
- **Misalignment of nearest neighbors:** In isomorphic language pairs, the nearest neighbors of a source word typically align well with those of its target counterpart. However, this is often not the case for non-isomorphic languages [73], such as English-Manipuri, English-Finnish, and English-Japanese. An example is shown in Figure 4.1, where the Manipuri word অমগা translates to the English word **additional**. While the five nearest neighbors of **additional**—**incentive, compensatory, ex-gratia, subsidy, budgetary**—are semantically related, the five nearest neighbors of অমগা—অমগগী (for someone); কয়াগা (how much); মশেল (themselves); মশেলগী (for themselves); মহৌশাগা (with nature)—do not share the same semantic relationships. This highlights the challenge of maintaining isomorphism in mapping-based methods, which assume that semantically similar word pairs should have similar local neighborhoods.

To address the limitations of frequency-based dictionary alignment, we propose a *graph centrality-aware ridge regression-based orthogonal mapping method*, which:

1. Improves the rank correlation between source and target dictionary word pairs in their respective monolingual corpora by leveraging graph centrality.
2. Mitigates the impact of misaligned word pairs by incorporating ridge regression to adjust the projection matrix estimation.

The proposed method selects seed dictionary pairs based on network centrality, estimates word pair comparability using centrality measures, and applies orthogonal mapping with

ridge regression. Our experimental evaluations on six state-of-the-art orthogonal frameworks across five language pairs—both isomorphic and non-isomorphic—demonstrate that the centrality-aware ridge regression method enhances performance across multiple tasks, including *Bilingual Dictionary Induction (BDI), Sentence Retrieval Task (SRT), and Machine Translation (MT)*. Furthermore, an ablation study highlights the effectiveness of centrality-based ridge regression in improving cross-lingual word embedding alignment.

4.1.1 Contribution

The main contributions of this study are:

- This Chapter highlights the importance of selecting seed dictionaries based on centrality measures rather than relying solely on word frequency in source and target dictionary pairs.
- It demonstrates that selecting dictionary pairs with similar centrality measures across source and target languages improves CLWE performance.
- An orthogonal ridge regression mapping is introduced to penalize dictionary pairs that exhibit lower comparability in centrality measures.

4.2 Related work

In mapping-based models, the projection matrix is generally derived using a bilingual seed dictionary between the source and target languages. [72] proposed a regression-based approach to learn a linear mapping function. An approximate linear mapping is trained to minimize orientation differences between the source and target language spaces. [74] introduced a regression approach with regularization. However, one notable observation from this study is that increasing regularization leads to a decline in bilingual dictionary induction accuracy. Furthermore, the paper highlighted that hubness, a phenomenon where some word vectors act as hubs appearing frequently as nearest neighbors, intensifies with regularization.

The concept of orthogonal projection as a normalization technique for word embeddings after transformation was first introduced by [26]. [3] proposed an exact solution using Singular Value Decomposition (SVD), making the computation feasible in linear time concerning vocabulary size. This approach is grounded in the assumption that imposing an orthogonal constraint preserves monolingual invariance, specifically maintaining word vector lengths

and angles. Building on this, [23] developed a weakly supervised self-learning method that reduces the need for bilingual dictionaries to a maximum of 25 entries. The process starts by learning a transformation matrix W from a small bilingual dictionary. This matrix W is then utilized to generate additional dictionary entries. It is assumed that the expanded dictionary improves upon the original seed dictionary, which is subsequently used to refine the transformation matrix W iteratively until convergence. This iterative method is widely known as VECMAP.[1]

A novel method that eliminates the necessity of expert bilingual signals was introduced by [75], which generates a *pseudo-dictionary* by identifying similar character strings appearing in both languages. This study demonstrated that to ensure consistency in the linear transformation, an orthogonality constraint must be enforced.

An empirical evaluation conducted by [76] reveals that even state-of-the-art unsupervised approaches struggle significantly when applied to linguistically distant languages. Their bilingual dictionary induction experiments, which spanned 15 diverse languages and 210 language pairs, showed that fully unsupervised methods remain ineffective for most language pairs. [12] argued that orthogonal mapping is only suitable for language pairs with inherently isomorphic embeddings. Their study found that such mapping performs suboptimally in supervised English-Japanese bilingual lexicon induction tasks. To address non-isomorphic language pairs, [12] introduced an Iterative Normalization technique that facilitates orthogonal alignment by ensuring that individual word vectors maintain unit length and zero mean centering in parallel.

A strategy for handling the isomorphic properties of embedding spaces was proposed by [73]. The study suggests that a bilingual dictionary consists of both near-isomorphic and non-isomorphic translation pairs. To enhance alignment performance, the paper introduced a method for selecting near-isomorphic dictionary pairs.

4.3 Methodology

Given a bilingual seed dictionary $D = \{x_i, z_i\}$ where i=1 to k, with k representing the dictionary size. Let X and Z be matrices consisting of the embeddings of the source and target words, respectively. As proposed by [72], a linear mapping between the source embedding $x_i \in X$ and target embedding $z_i \in Z$ can be formulated as an ordinary least squares regression (OLSR) problem as defined below.

[1] https://github.com/artetxem/vecmap

$$\min_{W} \sum_{i=1}^{k} \|\vec{x_i}W - \vec{z_i}\|^2 \tag{4.1}$$

Here W represents the projection matrix that maps the source language space X to the target language space Z. The matrix W can be optimized using stochastic gradient descent [77]. A challenge with this approach is that the projection matrix W may need to be orthogonal to maintain length normalization and monolingual invariance [26, 3]. Under this constraint, [3] proposed that W can be obtained as $W = VU^T$, where the column vectors of U and V are left-right singular vectors of Z^TX. This orthogonal projection approach has also been applied in [23, 42].

However, as noted by [12], this orthogonal projection performs poorly for non-isomorphic language pairs. The non-isometric nature of these language pairs negatively affects the quality of bilingual dictionaries constructed based on frequency occurrences, which are typically used in orthogonal projections. For non-isometric language pairs, the neighbors of a source word in one space may not align well with the neighbors of the corresponding target word. Moreover, [73] reported that selecting a high-quality seed dictionary from a large dictionary set is crucial for improving cross-lingual embeddings. Using a subset of well-selected dictionary pairs not only enhances alignment but also reduces computational overhead.

To mitigate the misalignment of dictionary word pairs that are selected based on frequency, we introduce a penalty term for less comparable dictionary word pairs. We propose ridge regression as follows:

$$\min_{W} \sum_{i=1}^{k} \|\vec{x_i}W - \vec{z_i}\|_2^2 + \lambda_i \|W\|_2^2 \tag{4.2}$$

Here, λ_i represents the penalty term for the i^{th} dictionary pair, derived using centrality measures of the source and target words in the dictionary pair. The process of estimating the penalty is discussed in Section 4.4.2. The optimal W can be derived in closed form as follows:

$$\begin{aligned} f(W) &= \min_{W} \textstyle\sum_{i=1}^{k} \|\vec{x_i}W - \vec{z_i}\|_2^2 + \lambda_i \|W\|_2^2 \\ &= \min_{W} \left\{(XW - Z)^T(XW - Z) + \lambda W^TW\right\} \end{aligned} \tag{4.3}$$

The matrix W that minimizes $f(W)$ can be computed as follows.

$$\begin{aligned} &\tfrac{\partial f(W)}{\partial W} = 0 \\ &\Rightarrow -2X^TZ + 2(X^TX + \lambda I)W = 0 \\ &\Rightarrow W = (X^TX + \lambda I)^{-1}X^TZ \end{aligned} \tag{4.4}$$

Where $\lambda = \sum_{i=1}^{k} \lambda_i$.To maintain length normalization and monolingual invariance, W must be orthogonal. The nearest orthogonal matrix $\hat{W}$ can be estimated using polar decomposition as defined in [78].

$$\begin{aligned}
\hat{W} &= W\sqrt{W^TW}^{-1} \\
&= U\Sigma V^T \left\{\sqrt{(U\Sigma V^T)^T U\Sigma V^T}\right\}^{-1} \\
&= U\Sigma V^T \left\{\sqrt{V\Sigma U^T U\Sigma V^T}\right\}^{-1} \\
&= U\Sigma V^T \left\{\sqrt{V\Sigma^2 V^T}\right\}^{-1} \\
&= U\Sigma V^T \frac{1}{\sqrt{V\Sigma^2 V^T}} \\
&= UV^T \frac{1}{\sqrt{I}} \\
&= UV^T
\end{aligned} \tag{4.5}$$

Where $W = U\Sigma V^T$ represents the singular value decomposition (SVD) of W, with U, V^T, and Σ are the left singular vector, right singular vector, and diagonal matrix of singular values arranged in descending order. The orthogonal matrix $\hat{W}$ is equivalent to the SVD of W, but with the diagonal matrix replaced by an identity matrix. This transformed matrix $\hat{W}$ projects the source vector space into the target space. Given a new source word x_i, its projection into the target space is computed as follows:

$$\vec{z_i} = \vec{x_i}\hat{W} \tag{4.6}$$

For all baseline methods discussed in Section 4.5, $\hat{W}$ will be utilized for linear orthogonal mapping.

4.4 Dataset

The majority of studies on cross-lingual embeddings in the literature have predominantly relied on parallel corpora. However, Manipuri, being a low-resource language, lacks a sufficiently large English-Manipuri parallel corpus. To address this limitation, this study utilizes a comparable corpus consisting of documents or articles. Comparable corpora do not contain direct translations but instead comprise monolingual corpora in different languages that describe the same or similar events. One such example is the Wikipedia[2] dataset.

[2] https://linguatools.org/tools/corpora/wikipedia-comparable-corpora/

Table 4.1: Statistics of data, **LP**: Language Pairs

LP	Platform	sentences		words		unique words	
En-Mn	**Sangai Express +Poknafam+PMI**	**En** 129,546	**Mn** 181,553	**En** 3.5M	**Mn** 3.3M	**En** 15,247	**Mn** 24,449
En-It	**European Parliament**	**En** 1.90 M	**It** 1.90M	**En** 49.6M	**It** 47.4M	**En** 151,017	**It** 219,976
En-Fi	**European Parliament**	**En** 1.92 M	**Fi** 1.92M	**En** 47.4M	**Fi** 32.2M	**En** 151,017	**Fi** 219,976
En-Hi	**CILT,IIT Bombay**	**En** 1.6M	**Hi** 1.6M	**En** 23.8M	**Hi** 24.6M	**En** 238,765	**Hi** 392,634
En-Ja	**opensubtitles.org kitsunekko.net d-addicts.com subscene.com**	**En** 2.8 M	**Ja** 2.8M	**En** 23.6M	**Ja** 21.5M	**En** 154,276	**Ja** 138,487

Table 4.2: Statistics of dictionary pairs, **LP**: Language Pairs

LP	**Global Pairs**	**After filtering** (D_{total})
En-Mn	5435	4200
En-It	115854	14797
En-Fi	43055	6599
En-Hi	38221	14830
En-Ja	35353	11502

The English-Manipuri comparable corpus used in this study is extracted from two prominent local online news platforms, Sangai Express[3] and Poknafam[4], through web crawling techniques. These platforms publish daily news articles in both English and Manipuri. Additionally, comparable data for English-Manipuri is collected from news updates posted on PMIndia[5] [79]. For English-Italian and English-Finnish language pairs, the Europarl[6] parallel corpus [54], derived from the European Parliament proceedings, is utilized. English-Hindi data is sourced from the dataset provided by the Centre for Indian Languages Technology, IIT Bombay [80]. Furthermore, a parallel subtitle corpus [81], derived from conversational dialogue, is employed for English-Japanese. Table 4.1 presents the statistical details of the experimental dataset used in this study.

Supervised cross-lingual approaches need bilingual dictionary pairs for cross-lingual alignment. Most of the supervised methods require a large number of dictionary pairs (5k dictionaries) [42] to achieve greater coverage. We consider an English-Manipuri bilingual dictionary obtained from Directorate of Language Planning and Implementation, Government of Manipur [7]. Manipuri, a low-resource language, has minimal bilingual dictionary pairs;

[3] https://www.thesangaiexpress.com/index.html
[4] http://www.poknapham.in
[5] https://www.pmindia.gov.in/en/
[6] https://www.statmt.org/europarl/
[7] https://www.dlpi.mn.gov.in/en/

moreover, the dictionary pair must also be present in the respective monolingual embedding mentioned in section 4.5. Transliterate pairs of loan words and Named Entities are used to solve this problem. Transliteration pairs are word pairs with the same phonetics characteristics and semantic meaning but a different written script. Moreover, transliterate pairs of loan words and Named Entity will likely have the same semantically similar neighbors. Transliterate pairs are extracted from the comparable corpus by a sequence-to-sequence [51] auto-encoder using grapheme-based representation [52] described in [53]. The transliterate pairs obtained above filter out pairs with target words that are not in the Manipuri corpus. Finally, transliteration pairs of loan words and Named-Entity are manually selected. For English-Italian, English-Finish, English-Hindi, and English-Japanese, the dictionary pairs published as a part of the MUSE[8] library are used for the experiment. The detailed statistics of dictionary pairs on all language pairs are given in Table 4.2.

4.4.1 Seed Dictionary Preparation

Given a global dictionary, D_{total} (provided in Table 4.2), our objective is to extract a seed dictionary $D \subset D_{total}$ based on a ranking score. As outlined in Section 4.1, prior research predominantly relies on frequency as the ranking criterion, wherein the highest-frequency words from the source corpus and their corresponding translations in the target corpus are selected. In contrast, this study introduces an alternative ranking approach based on graph centrality within a co-occurrence network. Specifically, we explore two centrality measures: a local proximity-based metric, namely degree centrality, and a global proximity-based metric, namely Eigenvector centrality, as elaborated in the following sections.

4.4.1.1 Graph Centrality

The Word co-occurrence (Bigram) graph G is built to evaluate the graph centrality of words. G represents an adjacency matrix corresponding to each monolingual corpus and is formally defined as follows:

$$G_{ij} = \begin{cases} 1, bigram(v_i, v_j) = True \\ 0, bigram(v_i, v_j) = False \end{cases} \tag{4.7}$$

In this context, the function $bigram(v_i, v_j)$ determines whether the nodes v_i and v_j form a bigram. In a bigram network, an edge exists between two words (nodes) if they appear consecutively in the corpus.

[8] https://github.com/facebookresearch/MUSE

Graph centrality quantifies the significance of a word in the corpus by leveraging the structural topology of the graph constructed from the corpus. This research article considers the following two centrality measures:

(i) *Degree Centrality*: The degree centrality of a node v is given by $C_D(v) = \frac{deg(v)}{\sum deg(u)}$, where $deg(v)$ denotes the degree of the node v (i.e., the number of edges in the undirected graph G).

(ii) *Eigenvector Centrality*: This measure assigns relative scores to all nodes in the network, based on the principle that connections to highly scored nodes contribute more to the score of a given node than connections to lower-scoring nodes. It is mathematically expressed as $x_v = \frac{1}{\lambda} \sum_{t \epsilon M(v)} x_t$, where $M(v)$ represents the set of neighbors of v, and λ is a constant.

4.4.1.2 Why centrality should be considered over frequency?

As discussed in Section 4.1, prior research on mapping-based CLWE has consistently used word frequency as the primary criterion for selecting a seed dictionary. The fundamental assumption behind this approach is that a frequently occurring word in the source language corpus will also have a high frequency in the target language corpus. While this assumption generally holds for isomorphic language pairs, it does not necessarily apply to non-isomorphic and linguistically distant pairs. To examine this claim, we analyze the corpora from two perspectives.

First, we compute the rank correlation coefficient between the rank scores of the most frequent words in the source corpus and the rank scores of their corresponding dictionary words in the target corpus. A high rank correlation coefficient would indicate that a frequently occurring word in the source corpus also retains its prominence in the target corpus. We compute this correlation for both frequency-based and centrality-based rankings and compare the resulting coefficients. Given that the En-Mn language pair contains the smallest bilingual dictionary, with only 4,200 word pairs, we ensure a fair comparison by considering the top 4,200 pairs across all five language pairs. The results, presented in Table 4.3, reveal that rank correlations obtained using centrality measures are consistently higher than those based on frequency. This finding suggests that constructing the seed dictionary using centrality-based rankings enhances its overall quality.

For this estimation, we rely on D_{total} (as shown in Table 4.2), which assumes a one-to-one (hard) mapping between source and target words. However, in reality, a single source word may correspond to multiple semantically similar target words. To account for such loosely

Table 4.3: Correlation analysis in dictionary pairs using frequency, Degree Centrality (DC) and Eigenvector Centrality (EC)

Language Pair	**frequency**	**DC**	**EC**
English-Italian	0.72	0.80	0.74
English-Hindi	0.74	0.76	0.75
English-Finnish	0.66	0.69	0.68
English-Japanese	0.43	0.49	0.44
English-Manipuri	0.15	0.28	0.45

associated relationships, we further conduct a semantic similarity evaluation of the top 20 words in both the source and target corpora, selecting these words based on both Frequency and Eigenvector centrality. In this evaluation, two words are considered similar if they convey closely related meanings, even if they are not explicitly present in the hard-mapped dictionary. This supplementary analysis reinforces the findings from the rank correlation study, even under a more relaxed evaluation framework. For evaluating semantic similarity, the WordNet of the respective languages is employed. WordNet is a lexical database curated by researchers and lexicographers, making it a highly reliable resource in natural language processing and computational linguistics. Due to its well-structured organization, the results provided by WordNet are widely considered accurate and trustworthy. The following WordNets are utilized for different language pairs: KangleiWordNet for En-Mn, ItalWordNet v.2 for En-It, FinnWordNet for En-Fi, Hindi WordNet for En-Hi, and Japanese WordNet for En-Ja.

WordNet generates multiple semantically similar words in the source language (English) corresponding to a given target word. From the top 20 selected words (based on either frequency or Eigenvector centrality measures) in both the source and target corpora, a total of $20 \times 20 = 400$ word pairs are generated for semantic similarity annotation. In these 400 comparison entries for each language pair, an entry is marked as **YES** if it matches the corresponding WordNet result, indicating correct semantic similarity, and as **NO** otherwise. The percentage of semantic similarity in the top 20 pairs is then computed based on the number of **YES** obtained. The results indicate that the percentage of semantic similarity is higher when using Eigenvector centrality compared to frequency-based selection. The detailed percentage values are provided in Table 4.4. given in Table 4.4.

4.4.2 Need for Ridge-Regression

Although selecting a seed dictionary based on centrality enhances rank correlation, the correlation remains relatively low for non-isomorphic languages compared to isomorphic languages, as shown in Table 4.3. This indicates a significant disparity in centrality measures

Table 4.4: Percentage of semantically similar words in 20 source and target words with highest centrality score and highest frequency across languages. **Frequency Vs EC (EigenVector Centrality**)

Language Pair	**EC**	**Frequency**
English-Italian	90%	65%
English-Hindi	70%	55%
English-Finnish	75%	55%
English-Manipuri	70%	55%
English-Japanese	60%	50%

between source and target words in the dictionary for non-isomorphic languages. To further enhance the quality of the seed dictionary, seed dictionary pairs are chosen such that their centrality measures are comparable and evaluated using the method proposed in [72] for the BDI task. Comparability is quantified using ratio measures, as defined in equation 4.8.

$$comparability(C_x, C_z) = \begin{cases} \frac{\left\{\frac{C_z}{C_x \times 100} \times C_x\right\} + \left\{\frac{C_z}{C_x \times 100} \times C_z\right\}}{2} if C_x > C_z \\ \frac{\left\{\frac{C_x}{C_z \times 100} \times C_x\right\} + \left\{\frac{C_x}{C_z \times 100} \times C_z\right\}}{2} if C_z > C_x \end{cases} \tag{4.8}$$

Here C_x and C_z represent the centrality scores of the source and target words, respectively. To ensure that dictionary pairs with both low and comparable centrality measures receive a lower rank, the average percentage of the ratio of source and target centrality scores is considered. Figure 4.2 illustrates improved performance in re-aligning dictionary pairs that exhibit comparable centrality measures. This serves as initial evidence that dictionary pairs with similar centrality measures contribute to better performance in BDI tasks.

Building upon the importance of comparable dictionary pairs, a model utilizing ridge regression instead of ordinary least squares regression (OLSR) is proposed for learning the linear mapping matrix W. To minimize the impact of dictionary pairs with the least comparable centrality measures in orthogonal linear mapping, a penalty term is introduced. This penalty is computed as the absolute difference between the centrality measures of the source and target words in the bilingual training dictionary, as defined below:

$$\lambda = \sum_{i=1}^{k} abs(C_{x_i} - C_{z_i}) \tag{4.9}$$

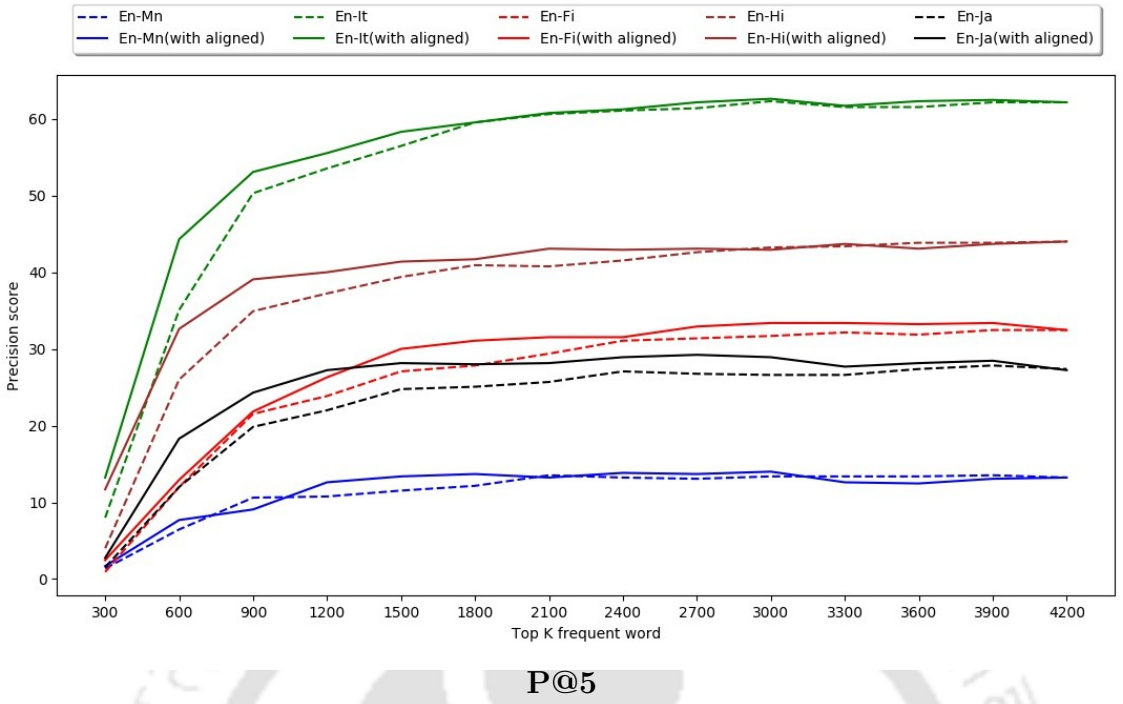


Figure 4.2: Comparison between non-aligned and comparable centrality aligned (degree centrality)

4.5 Experimental Setup

To conduct an empirical comparison with the proposed approach, three supervised methods [72, 3, 12], along with one weakly supervised method [23] utilizing 25 dictionary pairs, are selected as baseline models. These models are widely regarded as standard baselines in numerous studies related to orthogonal linear mapping. Additionally, the weakly supervised method [23] is also assessed in a supervised setting with self-learning. The iterative normalization approach [12] is evaluated under both the VECMAP and MUSE settings. A detailed overview of these methods is provided in Table 4.5.

For this study, monolingual word embeddings are generated using Word2Vec with the Continuous Bag of Words (CBOW) [82] architecture, employing a dimensionality of 300, a window size of 5, and filtering out words with a frequency count below ten. Furthermore, both the baseline and the proposed methods are evaluated using pre-trained Multilingual BERT for language pairs En-It, En-Fi, En-Hi, and En-Ja. For En-Mn, the pre-trained embeddings extracted from IndicBERTv2-MLM-only are utilized, as Manipuri data is not included in mBERT. The embedding dimensionality for both Multilingual BERT and IndicBERTv2-MLM-only is set to 768.

4.5.1 Bilingual Dictionary Induction

The training dictionary is initially partitioned into the top-kk pairs based on frequency, where kk varies between 300 and 4,200 across all language pairs. To ensure fair evaluation, 50 pairs from each top-kk subset are designated as testing pairs, maintaining mutual exclusivity between training and testing sets. This process results in a total of 700 testing pairs across all language pairs. The remaining top-kk training pairs, excluding the selected testing pairs, are then reorganized based on centrality to form centrality-based top-kk training pairs. Notably, the same set of testing pairs is utilized in both frequency-based and centrality-based settings.

4.5.2 Cross-lingual Sentence Retrival Task (CSRT)

The proposed and baseline models are also assessed using a sentence-level intrinsic task: Cross-lingual Sentence Retrieval Task (CSRT). We evaluate our proposed model on cross-lingual embeddings generated by [3], utilizing the top-300 dictionary pairs obtained from monolingual CBOW embeddings. These dictionary pairs are derived based on three alignment strategies: Frequency Align, Degree Centrality Align, and Eigenvector Centrality Align. For each language pair, we use 10,000 parallel sentences, with the first 1,000 sentences designated for testing. Specifically for En-Mn, the parallel sentences are sourced from the PMI corpus, whereas for En-It, En-Fi, En-Hi, and En-Ja, they are extracted from the dataset described in Section 4.4.

4.5.3 Machine Translation

Subsequently, the proposed model and the baseline model are assessed on an extrinsic sequence-to-sequence task: machine translation. We utilize the Fairseq sequence modeling toolkit [83] for this evaluation. Similar to the CSRT evaluation, our proposed model is tested using the cross-lingual embeddings generated by [3], constructed with the top-300 dictionary pairs derived from monolingual CBOW embeddings. These dictionary pairs are obtained through three distinct alignment methods: Frequency Align, Degree Centrality Align, and Eigenvector Centrality Align. Instead of initializing the Fairseq model with random embeddings, we incorporate the aforementioned cross-lingual embeddings. The embedding dimension provided to Fairseq is set to 300. For training, we consider 100K parallel sentences, along with 1000 parallel sentences each for validation and testing across all five language pairs. The extraction of these parallel sentences follows the procedure described in Section 4.5.2.

Table 4.5: Details of the baseline methods

	paper	method
supervised	Mikolov et al. (2013a)[72]	Mapping (Regression)
	Artetxe et al. (2016)[3]	Orthogonal Mapping
	Artetxe et al. (2017)[23]	Orthogonal Mapping
	Zhang et al. (2019)[12] +VECMAP	iterative normalization
	Zhang et al. (2019)[12] +MUSE	iterative normalization
semi-supervised	Artetxe et al. (2017)[23] (25 pairs)	Orthogonal Mapping (refinement)

4.5.4 Evaluation metrics

The baseline and proposed methods are assessed on two tasks: Bilingual Dictionary Induction (BDI) and Cross-lingual Sentence Retrieval Task (CSRT), using P@1 (Precision at 1) and P@5 (Precision at 5) as evaluation metrics. P@1 accounts for only the first nearest neighbor, whereas P@5 considers up to the 5^{th} nearest neighbor when compared against the ground truth. Since there is a one-to-one correspondence in bilingual dictionary pairs across all five language pairs, P@1 and P@5 effectively measure the percentage of exact word translation matches found within the first and top five nearest neighbors, respectively, in the given test set. For machine translation, the evaluation metric used is the BLEU-4 score, which specifically computes precision based on 4-gram (four consecutive words) matches. This metric evaluates not just individual word matches but also sequences of four words within the translation output.

4.6 Results and discussion

Table 4.3 indicates that language pairs with close linguistic relationships exhibit a higher correlation between the centrality measures of source and target words in the dictionary. This implies that dictionary pairs in closely related languages tend to have comparable centrality measures, indirectly suggesting that these languages possess a more isomorphic structure compared to distant language pairs. Furthermore, Figure 4.2 demonstrates that language pairs with higher correlation yield improved performance in BDI tasks. It is also evident from Figure 4.2 that dictionary pairs with comparable centrality measures achieve better performance in BDI at P@5 compared to aligning dictionary pairs without considering comparability. This finding motivates the introduction of a penalty for dictionary pairs with the least comparable centrality measures.

4.6.1 BDI results on CBOW embeddings

The evaluation outcomes of the proposed approach at top-300 across five baseline methods, as detailed in Section 4.5, are presented in Table 4.6. Additionally, the Table includes the performance of the proposed method using 25 dictionary pairs compared to the baseline approach [23]. From Table 4.6, it can be seen that our proposed model surpassed the baseline models in 91.67%, 93.75%, 95.83%, 91.67%, and 87.50% of cases for En-Mn, En-It, En-Fi, En-Hi, and En-Ja, respectively, in the BDI task.

Table 4.6: Result of evaluation on top-300 and top-25 in five languages pairs using CBOW embeddings. Bold represent highest P@1 and P@5. FA: Frequency Align (Baseline), DCA: Degree Centrality Align, DCAR: Degree Centrality Align and Regularised, ECA: Eigen Centrality Align, ECAR: Eigenvector Centrality Align and Regularised.

LP	Supervised Method	P@1					P@5				
		FA	DCA	DCAR	ECA	ECAR	FA	DCA	DCAR	ECA	ECAR
En-Mn	Mikolov (2013a)[72]	00.15	01.15	**02.15**	00.46	01.38	01.38	01.23	**05.38**	01.38	04.61
	Artetxe (2016)[3]	04.14	**05.85**	**05.85**	05.14	05.28	12.14	13.86	13.98	13.14	**14.13**
	Artetxe (2017)[23]	08.14	08.28	**09.57**	08.86	08.86	18.42	18.43	19.57	19.57	**20.57**
	Zhang (2019)[12]+MUSE	05.74	**07.86**	07.71	07.43	07.28	12.95	15.28	15.88	06.00	**16.44**
	Zhang (2019)[12]+VECMAP	10.00	10.43	**10.71**	09.43	**10.71**	19.00	19.57	**19.86**	**19.86**	**19.86**
	Artetxe (2017)[23]25 pairs	06.71	07.57	**09.14**	07.14	08.00	14.57	17.57	**19.57**	17.28	18.28
En-It	Mikolov (2013a)[72]	00.77	01.08	**03.85**	00.61	02.61	03.85	04.15	**11.23**	02.61	06.77
	Artetxe (2016)[3]	25.28	29.00	**29.71**	27.71	27.71	45.71	**48.00**	45.14	47.28	47.28
	Artetxe (2017)[23]	37.57	42.28	43.28	43.00	**45.00**	54.00	59.00	59.57	59.28	**60.71**
	Zhang (2019)[12]+MUSE	19.57	39.28	**39.71**	37.86	38.57	37.42	56.42	**56.43**	55.86	55.14
	Zhang (2019)[12]+VECMAP	39.71	42.43	44.57	43.71	**45.14**	57.43	59.14	60.57	59.86	**62.14**
	Artetxe (2017)[23]25 pairs	29.29	38.57	42.43	40.71	43.71	49.71	56.71	58.57	56.00	**62.57**
En-Fi	Mikolov (2013a)[72]	00.46	00.61	**02.77**	00.31	01.38	00.92	01.85	**08.61**	01.23	04.92
	Artetxe (2016)[3]	05.29	06.00	06.14	**06.28**	05.57	16.00	17.71	17.86	17.28	**18.14**
	Artetxe (2017)[23]	14.57	17.00	**18.28**	17.86	18.00	26.71	29.00	**31.00**	29.43	**31.00**
	Zhang (2019)[12]+MUSE	00.00	02.86	02.93	**03.00**	**03.00**	00.86	05.28	05.28	**08.43**	**08.43**
	Zhang (2019)[12]+VECMAP	14.28	17.10	17.86	17.57	**18.00**	28.57	28.14	29.29	29.57	**31.00**
	Artetxe (2017)[23]25 pairs	01.86	02.00	02.00	03.29	**03.78**	04.00	06.00	**06.67**	06.14	06.57
En-Hi	Mikolov (2013a)[72]	00.46	00.77	**01.38**	00.61	01.08	02.00	01.38	**05.69**	01.23	03.38
	Artetxe (2016)[3]	13.43	**15.43**	14.57	13.43	13.86	27.00	30.14	30.71	26.71	**31.00**
	Artetxe (2017)[23]	24.43	26.57	26.43	**26.71**	25.86	39.71	42.00	**43.28**	43.14	43.14
	Zhang (2019)[12]+MUSE	15.71	15.86	16.57	**17.43**	**17.43**	30.43	31.43	**31.74**	30.57	30.89
	Zhang (2019)[12]+VECMAP	22.85	**28.28**	**28.28**	28.14	28.14	40.57	42.71	**43.14**	**43.14**	42.86
	Artetxe (2017)[23]25 pairs	23.71	24.71	24.86	25.00	**25.28**	38.00	38.86	38.86	**40.86**	**40.86**
En-Ja	Mikolov (2013a)[72]	00.61	00.92	**03.38**	000.77	03.08	01.23	01.38	02.00	03.08	**08.69**
	Artetxe (2016)[3]	05.71	06.28	06.28	**06.43**	06.28	14.71	15.00	15.56	**15.57**	**15.57**
	Artetxe (2017)[23]	09.43	11.14	**12.28**	11.71	11.86	21.00	22.28	23.86	24.14	**25.00**
	Zhang (2019)[12]+MUSE	00.86	07.43	07.86	**09.00**	**09.00**	02.57	14.57	15.86	16.28	**17.00**
	Zhang (2019)[12]+VECMAP	11.14	13.00	13.45	**14.00**	13.00	21.71	**24.71**	**24.71**	24.00	24.43
	Artetxe (2017)[23]25 pairs	04.43	03.57	01.57	02.14	**04.57**	10.00	09.71	04.43	08.43	**11.71**

In En-Mn, the score 91.67% indicates that out of 48 cases, the proposed model outperformed the baseline models in 44 instances. Since each baseline model accounts for 8 cases, considering both P@1 and P@5, this results in a total of 48 cases per language pair across

the six baseline methods, as presented in Table 4.6. A notable performance improvement of 5% and 12.86% is observed in comparison to the iterative refinement baseline model [23] when utilizing 25 dictionary pairs in En-Mn and En-It, respectively. This underscores the effectiveness of the proposed model, particularly when using 25 dictionary pairs, signifying a substantial enhancement. For En-Ja, the proposed model, which integrates iterative refinement and employs 25 dictionaries, demonstrates performance gains, especially when Eigenvector centrality is considered. However, the improvement in Eigenvector centrality is relatively less pronounced compared to other language pairs. A comprehensive evaluation of top-k dictionary pairs, where k=300k=300 to 4,2004,200, is detailed in **??**. The results indicate that the proposed method consistently outperforms the baseline model in most cases. These findings highlight that penalizing dictionary pairs with the lowest comparable centrality scores using ridge regression leads to improved performance in BDI tasks.

4.6.2 BDI results on top-300 to 4,200

In this section, the evaluations ranging from top-300 to top-4,200 are illustrated in Figures 4.3, 4.4, and 4.5. This evaluation is conducted on 700 test pairs, as described in Section 4.5, at both P@1 and P@5. The results indicate that incorporating regularization enhances performance in the BDI task **78.57%**, **81.07%**, **84.28%**, **83.57%**, and **93.21%** of the time for **En-Mn**, **En-It**, **En-Fi**, **En-Hi**, and **En-Ja**, respectively. Specifically, **78.57%** indicates that, out of 280 instances, centrality-based regularization enhances 220 cases in **En-Mn**. Likewise, **81.07%**, **84.28%**, **83.57%**, and **93.21%** signify that, out of 280 cases, regularization with centrality improves 227, 236, 234, and 261 instances in **En-It**, **En-Fi**, **En-Hi**, and **En-Ja**, respectively, in the BDI task.

The evaluation in [72] for top-300 to top-4,200 at P@1 results in 14 instances, and the same assessment at P@5 also yields 14 instances, totaling 28+28=56 cases when evaluating [72] using **Degree Centrality** and **Eigenvector Centrality**. Similarly, for each of the methods proposed in [3], [23], [12]+MUSE, and [12]+VECMAP, an additional 56 cases are considered. Consequently, a total of 56x5=280 cases are evaluated across **five** different methods. The findings demonstrate that aligning dictionary pairs using centrality information while penalizing those with lower comparable measures significantly enhances BDI task performance.

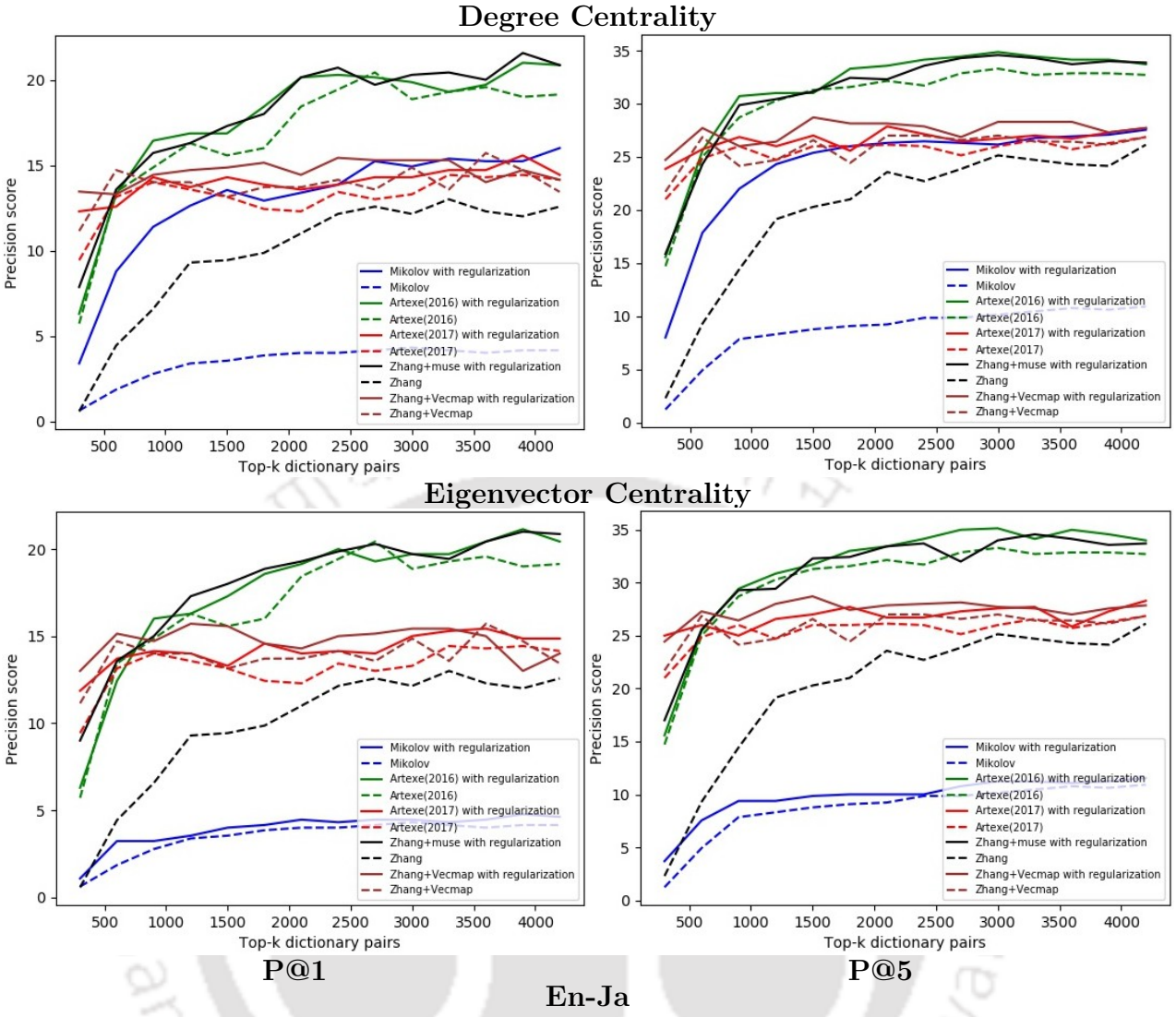


Figure 4.3: Evaluation on top-300 to 4,200 for En-Ja

4.6.3 BDI results on mBERT embeddings

Examining the proposed method within the context of Large Language Model (LLM) embeddings is crucial for evaluating its effectiveness in real-world linguistic scenarios, thereby ensuring its applicability beyond controlled experimental settings. As observed in Table 4.7, our proposed method outperforms the baseline model (Frequency align) when utilizing mBERT embeddings. However, its performance remains significantly lower than that of Word2Vec (CBOW). In certain cases, both the baseline and the proposed methods yield

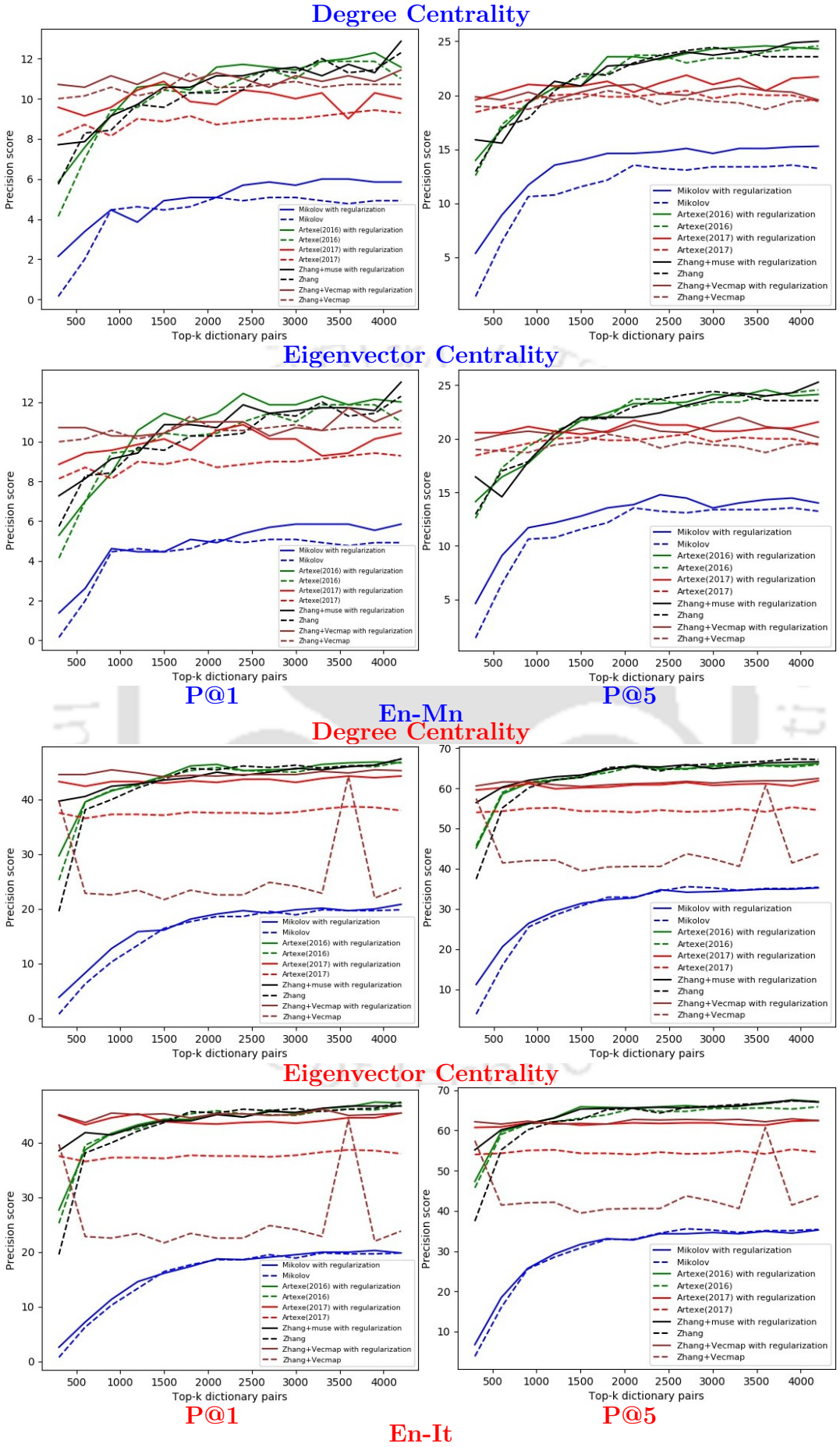


Figure 4.4: Evaluation on top-300 to 4,200 for En-Mn and En-It

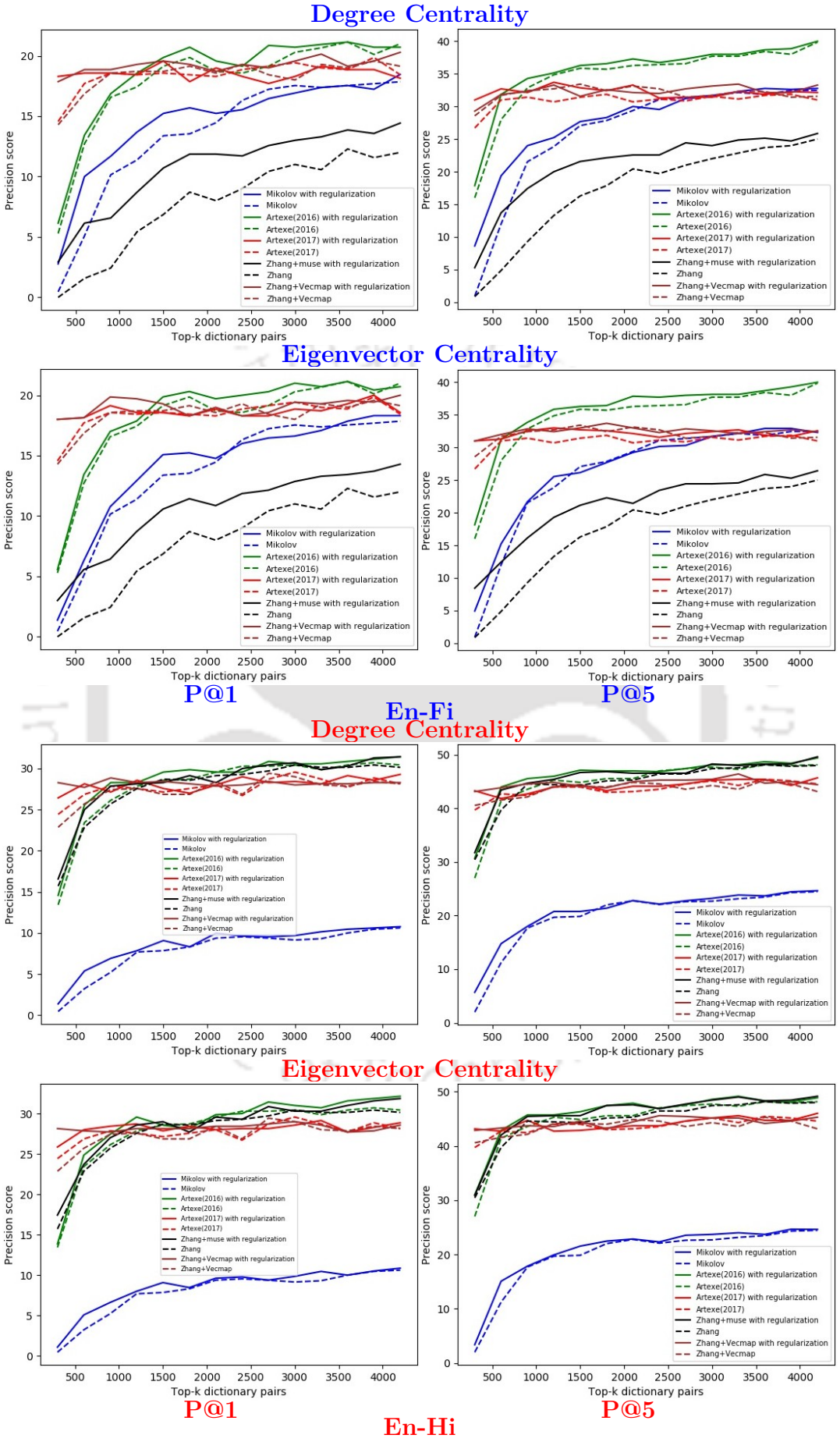


Figure 4.5: Evaluation on top-300 to 4,200 for En-Fi and En-Hi

Table 4.7: Result of evaluation on top-300 and top-25 dictionary in En-It, En-Fi, En-Hi, En-Ja languages pairs using MBert embeddings and En-Mn using IndicBERTv2-MLM-only. Bold represent highest P@1 and P@5. FA: Frequency Align (Baseline), DCA: Degree Centrality Align, DCAR: Degree Centrality Align and Regularised, ECA: Eigenvector Centrality Align, ECAR: Eigenvector Centrality Align and Regularised

LP	Supervised Method	P@1					P@5				
		FA	DCA	DCAR	ECA	ECAR	FA	DCA	DCAR	ECA	ECAR
En-Mn	Mikolov (2013a)[72]	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
	Artetxe (2016)[3]	00.00	00.00	00.00	00.00	00.10	00.00	00.14	**00.28**	00.14	**00.28**
	Artetxe (2017)[23]	00.00	00.00	00.00	00.00	00.00	00.00	00.14	**00.28**	00.14	**00.28**
	Zhang (2019)[12]+MUSE	00.00	00.14	**00.71**	00.14	**00.71**	00.00	00.14	**01.71**	00.28	**01.71**
	Zhang (2019)[12]+VECMAP	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
	Artetxe (2017)[23]25 pairs	00.00	00.00	00.00	00.00	00.00	00.00	00.14	**00.14**	00.00	00.00
En-It	Mikolov (2013a)[72]	00.00	00.00	00.00	00.00	00.00	00.00	00.00	**01.00**	00.00	**01.00**
	Artetxe (2016)[3]	09.71	**10.00**	**10.00**	09.71	09.71	19.00	20.00	**20.57**	09.28	19.71
	Artetxe (2017)[23]	20.28	20.28	20.28	20.28	20.28	31.14	31.28	31.57	31.28	**31.78**
	Zhang (2019)[12]+MUSE	01.86	02.43	02.57	02.14	02.71	04.43	04.86	04.22	04.86	**6.28**
	Zhang (2019)[12]+VECMAP	20.28	20.57	**20.86**	20.71	20.71	32.14	32.43	**32.71**	32.29	**32.71**
	Artetxe (2017)[23]25 pairs	19.14	19.86	**20.00**	19.86	**20.00**	30.71	30.57	**31.29**	30.57	**31.29**
En-Fi	Mikolov (2013a)[72]	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
	Artetxe (2016)[3]	00.14	00.28	**01.28**	00.28	00.28	01.28	01.38	**02.14**	01.38	02.00
	Artetxe (2017)[23]	01.57	01.86	**02.71**	01.71	01.71	03.29	03.43	03.57	03.43	**04.00**
	Zhang (2019)[12]+MUSE	00.00	00.00	**00.57**	00.00	00.28	00.00	00.00	**1.14**	00.00	00.43
	Zhang (2019)[12]+VECMAP	02.14	**02.86**	02.29	02.71	02.83	03.86	04.29	**04.86**	04.29	**04.86**
	Artetxe (2017)[23]25 pairs	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00	00.00
En-Hi	Mikolov (2013a)[72]	00.00	00.00	**00.43**	00.00	**00.43**	00.00	00.00	**00.43**	00.00	**00.43**
	Artetxe (2016)[3]	00.43	00.71	01.00	01.00	**01.86**	01.43	02.14	02.28	01.71	**02.71**
	Artetxe (2017)[23]	00.29	00.71	01.00	00.57	**00.86**	02.43	02.86	**03.71**	02.86	03.57
	Zhang (2019)[12]+MUSE	00.28	00.43	**00.86**	00.38	**00.86**	00.57	01.14	**02.43**	01.43	01.86
	Zhang (2019)[12]+VECMAP	00.86	**01.71**	**01.71**	01.00	01.14	03.43	03.83	03.86	03.83	**04.29**
	Artetxe (2017)[23]25 pairs	00.00	00.00	00.00	00.00	00.00	00.00	00.14	**00.29**	00.00	00.14
En-Ja	Mikolov (2013a)[72]	00.00	00.00	00.00	00.00	**00.14**	00.00	00.00	00.00	00.00	**01.00**
	Artetxe (2016)[3]	01.00	**01.86**	**01.86**	01.71	**01.86**	02.71	02.86	03.71	03.43	**04.00**
	Artetxe (2017)[23]	04.14	04.71	04.71	04.28	**05.00**	07.14	07.54	07.83	07.71	**08.71**
	Zhang (2019)[12]+MUSE	00.00	00.28	00.71	00.00	**00.86**	00.00	00.71	**02.14**	00.71	**02.14**
	Zhang (2019)[12]+VECMAP	02.43	04.00	**04.71**	04.14	04.57	04.86	06.43	07.29	07.71	**08.86**
	Artetxe (2017)[23]25 dictionary pairs	00.00	00.14	**00.29**	00.14	**00.29**	00.00	00.29	**00.43**	00.29	**00.43**

zero performance, such as when applying the approach introduced in [8] to En-Mn. This outcome may be attributed to the inadequate lexical information captured by mBERT embeddings. Since mBERT assigns embeddings to word tokens based on their surrounding context and positional information within a sentence, its representation is inherently influenced by these factors. Additionally, the process of splitting words into sub-words is likely to degrade lexical information, as noted in [70].

Table 4.8: Result of evaluation of Sentence Retrieval Task (SRT) on top-300 dictionary pairs over [3] in five languages pairs using CBOW embeddings. Bold represent highest score. FA: Frequency Align (Baseline), DCAR: Degree Centrality Align and Regularised, ECAR: Eigenvector Centrality Align and Regularised

LP	P@1			P@5		
	FA	DCAR	ECAR	FA	DCAR	ECAR
En-Mn	02.90	03.00	**03.30**	06.60	06.40	**06.90**
En-It	05.30	**06.60**	06.10	13.70	**14.40**	14.00
En-Fi	03.70	**04.00**	03.70	08.30	08.10	**08.80**
En-Hi	00.60	0.70	**0.90**	01.70	02.10	**03.00**
En-Ja	00.50	00.70	**00.80**	01.30	01.60	**01.90**

Table 4.9: Result of evaluation of Machine Translation on top-**300** dictionary pairs over [3] in five languages pairs using CBOW embeddings. Bold represent highest BLUE4 score. FA: Frequency Align (Baseline), DCAR: Degree Centrality Align and Regularised, ECAR: Eigenvector Centrality Align and Regularised

LP	FA	DCAR	ECAR
En-Mn	09.45	19.52	**19.86**
En-It	23.79	**24.31**	24.14
En-Fi	18.59	18.71	**18.75**
En-Hi	18.61	19.40	**19.61**
En-Ja	06.17	**06.40**	06.19

4.6.4 CSRT and Machine Translation

Evaluating the proposed model on sentence-level intrinsic tasks, such as Cross-lingual Sentence Retrieval Tasks (CSRT), is essential for assessing its ability to accurately retrieve relevant information, which directly impacts applications like information retrieval systems. Additionally, analyzing its performance in extrinsic tasks, such as Machine Translation (MT), provides valuable insights into its practical applicability. As shown in Table 4.8, the proposed method demonstrates superior performance over the baseline in SRT, with the most significant improvement observed in En-Hi, where Eigenvector Centrality Align and Regularised (ECAR) achieves a 1.30% increase. Similarly, Table 4.9 indicates that the proposed method outperforms the baseline in MT, with the highest gain in BLEU4 score observed in En-Hi, recording a 1% increase under ECAR. These results confirm that centrality-aware cross-lingual mapping enhances not only Bilingual Dictionary Induction (BDI) but also other intrinsic tasks (e.g., SRT) and extrinsic tasks (e.g., MT).

Table 4.10: Evaluation of cross-lingual intrinsic word embedding vector algebra. FA: Frequency Align (Baseline), ECAR: Eigenvector Centrality Align and Regularised. The English meaning words are given in the next row.

English word	FA	ECAR
tamil+nadu	ছত্তিসগর, গুজরাত, কেরলা, মহারাস্ত্র, •	**কর্নাতকা**, **কেরলা**, **অন্ধ্র**, ছত্তিসগর, ত্রিপুরা
	chhattisgarh, Gujarat, Kerala, Maharastra, dot	Karnataka, Kerala, Andhra, chhattisgarh, Tripura
Kingfisher-fisher	হাওখা, আশাশিংনা, চাক্কৌবগা, তরুনি, অকায়ব	অলিগর, এন্নুয়েল, ভোটিং, **ইরাবোত**, **ইরাবতপু**
	place name, hopes, feast, name , broken	place name, annual, voting, irabot, irabot,
tamil-karnataka+kerala	কেরলা, ত্রিপুরা, তামিল, মহারাস্ত্র, নাদুদা	আসামদা, কেরলা, তামিল, ত্রিপুরা, **পুদুচেরিগী**
	kerala, tripura, tamil, maharastra, at nadu	at assam, kerala, tamil, tripura, puducherry

4.6.5 Word Embeddings Vector Algebra operation

Evaluating the intrinsic vector algebra of word embeddings is essential for analyzing the semantic consistency and mathematical properties of cross-lingual embeddings. This assessment ensures that word relationships remain intact across different languages, enabling meaningful cross-lingual comparisons. By validating the capability of cross-lingual embeddings to capture linguistic subtleties and cultural context, this evaluation enhances their effectiveness in multilingual natural language processing tasks. For this evaluation, we consider three distinct operations, as outlined below:

(i) word1+word2

(ii) word1-word2

(iii) word1-word2+word3

In Table 4.10, it is observed that the proposed method, **ECAR**, successfully places the three southern states—Karnataka, Kerala, and Andhra—among the top-3 nearest neighbors of Tamil+Nadu. In contrast, the baseline method, **FA**, fails to achieve this, as it ranks northern states like Chhattisgarh and Gujarat as the top-2 nearest neighbors instead. The case of Kingfisher-fisher presents an intriguing observation, where the proposed method includes **Irabot** among the five nearest neighbors. Irabot was a renowned leader and social reformer from Manipur, India. This demonstrates that the proposed method effectively captures the semantic essence of **king** from kingfisher-fisher within a cross-lingual framework. Furthermore, the proposed method successfully identifies **Puducherry** from tamil-karnataka+kerala in a cross-lingual setting, a task that the baseline method is unable to accomplish.

Table 4.11: Average Hub score and Performance in BDI (Regularised Vs Non-Regularised) and **LP**: Language pairs, **bold** represent higher score and ↑ represent an increase in Hub score.

LP	**Non-Regularised**			**Regularised**		
	Average hub score					
	k=1	k=5	k=20	k=1	k=5	k=20
En-Mn	01.39±01.51↑	01.91±02.92	02.95±05.49	01.39±01.48	01.96±03.03↑	03.06±05.85↑
En-It	01.11±00.46	01.58±01.23	02.65±02.77	01.11±00.46	01.58±01.23	02.65±02.77
En-Fi	01.26±01.06↑	01.65±01.95↑	02.50±03.66↑	01.21±00.72	01.59±01.45	02.42±02.99
En-Hi	01.25±00.67	01.90±01.67	03.36±03.86	01.25±00.69	01.91±01.71↑	03.38±03.92↑
En-Ja	01.76±02.11	02.91±04.60	05.22±09.11	01.82±02.33↑	03.01±04.77↑	05.38±09.51↑
LP	**Performance**					
	P@1	P@5	P@20	P@1	P@5	P@20
En-Mn	05.40	11.80	19.80	**06.93**	**13.07**	**21.40**
En-It	37.10	53.07	63.53	**37.40**	**53.27**	**63.80**
En-Fi	11.13	20.53	29.67	**11.33**	**21.02**	**30.53**
En-Hi	21.60	37.13	**50.33**	**50.33**	37.13	50.20
En-Ja	05.27	12.07	21.00	**05.47**	**12.67**	**21.40**

4.6.6 Does regularization increase hubness?

The regularization technique introduced in [74] demonstrates that applying regularization markedly increases hubness, thereby deteriorating the performance of the BDI task. Hubness, a well-documented phenomenon in high-dimensional spaces [84], occurs when a target word frequently becomes the nearest neighbor of multiple source words. The hub score of a target word is determined by the number of times it appears among the nearest-k neighbors of distinct source words, where k=1,5,20. A higher hub score implies a greater likelihood of the target word being mapped to an unrelated source word, ultimately impairing the effectiveness of the BDI task. The average hub score and standard deviation are computed for both regularized and non-regularized cases. Using the same number of dictionary pairs as in [74], the top 6,500 centrality score-aligned dictionary pairs for En-It, En-Fi, En-Hi, and En-Ja are selected. These 6,500 pairs are split equally into two halves, with 3,250 pairs allocated for a balanced evaluation. From each half, the top 2,500 pairs are designated as training data, resulting in a total of 5,000 training pairs. The remaining 750 pairs from each half are reserved for testing, leading to a total of 1,500 testing pairs. In the case of En-Mn, due to the limited availability of dictionary pairs, 4,341 pairs are utilized. Applying the same partitioning approach, 2,841 pairs are used for training, while 1,500 pairs are assigned for testing. The evaluation of BDI task performance follows the approach introduced in [72], assessing both regularized and non-regularized methods at P@1, P@5, and P@20, to investigate the correlation between hub score and BDI performance.

Hubness analysis from Table 4.11 reveals that the average hub score remains constant, with only a negligible increase under the regularized setting. The most significant increase is observed in the En-Ja case at k=20k=20, reaching a value of 0.16. Conversely, in the En-Fi scenario, regularization leads to a decrease in hubness. Furthermore, a slight rise in the average score does not negatively impact performance in BDI tasks; instead, an

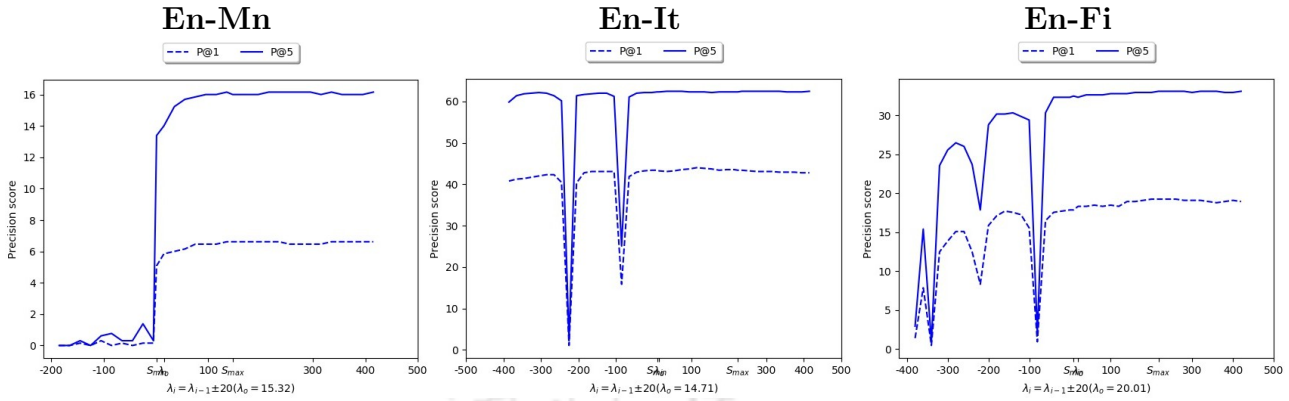
improvement is noted. This suggests that regularization does not inherently amplify hubness or degrade performance in BDI tasks. The proposed centrality-based ridge-regression model incorporates a regularization term ( ) as part of the loss function to penalize dictionary pairs with the least comparable centrality measures, distinguishing it from the used in [74].

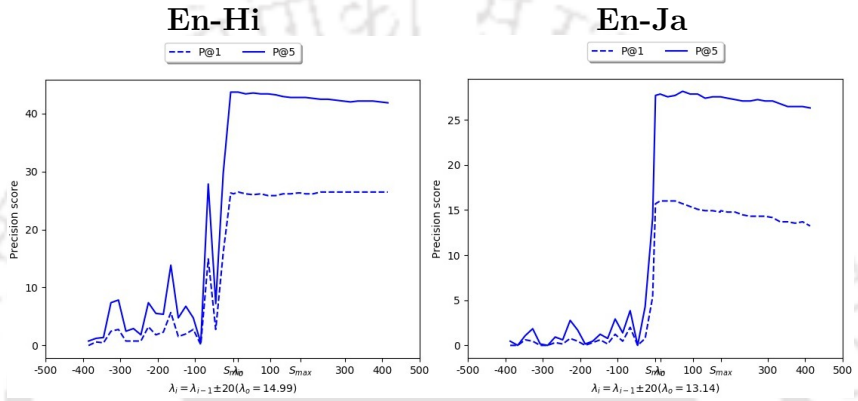
4.6.7 Is Graph Centrality based λ good reference point and stable?

The regularization parameter (λ), derived from centrality measures as detailed in Section 4.4.2, is modified by adding $\pm\delta$, i.e., $\lambda \pm \delta$, and assessed using the approach outlined in [72] and [12]+MUSE for the BDI task at top-300 and top-4,200 ranks. The value of δ is set to 10 and 20 for top-300 and top-4,200, respectively. This variation in λ aims to analyze its impact on BDI task performance.

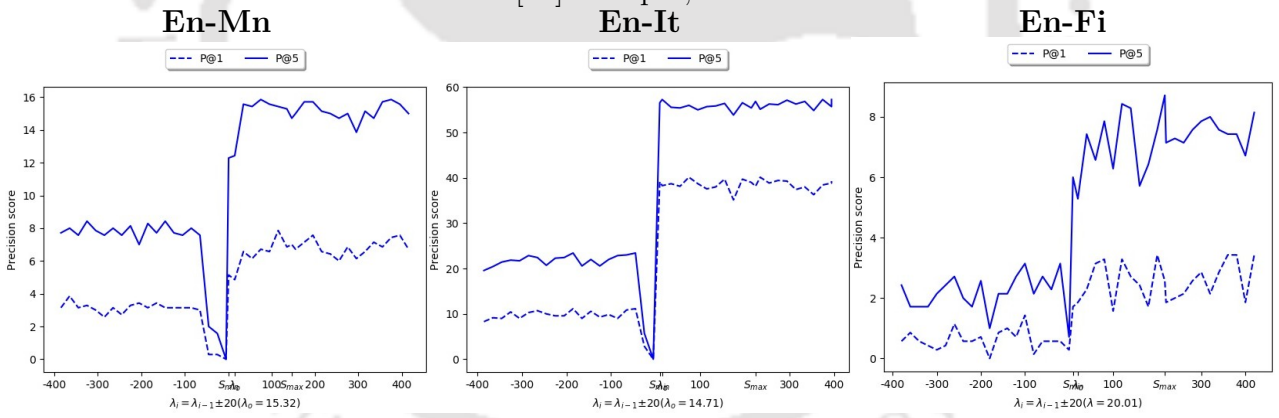
The analysis of λ variation reveals that the value computed from centrality measures remains stable and serves as a reliable reference. Increasing λ ensures stability in P@1 and P@5 at top-4,200. In contrast, decreasing λ initially maintains the performance but eventually results in a sharp decline after crossing the minimum singular value (S_{min}) of the source language embedding matrix (X), as evaluated using the method in [72] and [12]+MUSE. Notably, a minor improvement is observed before the significant drop in performance. These findings indicate that the λ value obtained from centrality measures is consistent, non-random, and a valid reference point. An interesting observation is that λ is always greater than the minimum singular value (S_{min}) of the source language embedding matrix (X). A detailed visualization of this analysis, evaluated using Eigenvector centrality, is presented in Figure 4.6.

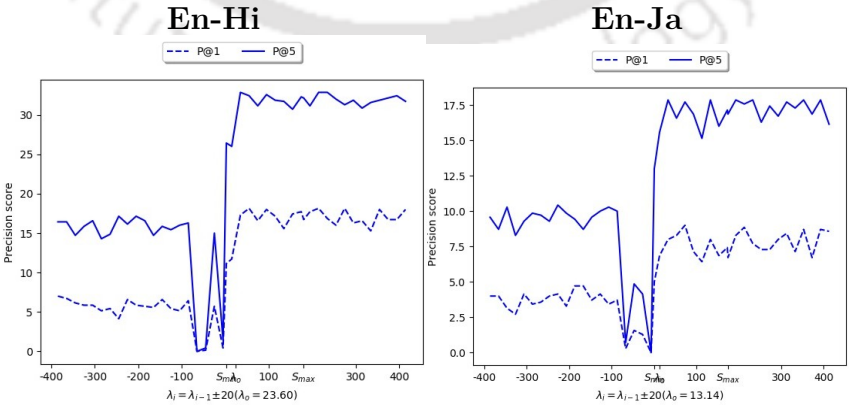
A similar evaluation is conducted using degree centrality, as illustrated in Figures 4.7 and 4.8, to further assess the reliability of λ. When λ increases, P@1 and P@5 performance improves for top-300, with the rate of improvement gradually decreasing as λ continues to rise. For top-4,200, performance remains stable as λ increases when evaluated using [72] and [12]+MUSE. However, for En-Hi and En-Ja, an increase in λ leads to a decline in P@1 and P@5 performance at top-4,200 when assessed using [72]. The evaluation based on **Eigenvector Centrality** for top-300 is presented in Figure 4.9. Similarly, it is observed that the λ value derived from centrality measures remains a stable and meaningful reference, consistently exceeding the minimum singular value (S_{min}) of the source language embedding matrix X.





[72] at top-4,200





[12]+MUSE at top-4,200

Figure 4.6: Evaluation on varying $\lambda_i = \lambda_{i-1} \pm \delta$ (20) using Eigenvector Centrality

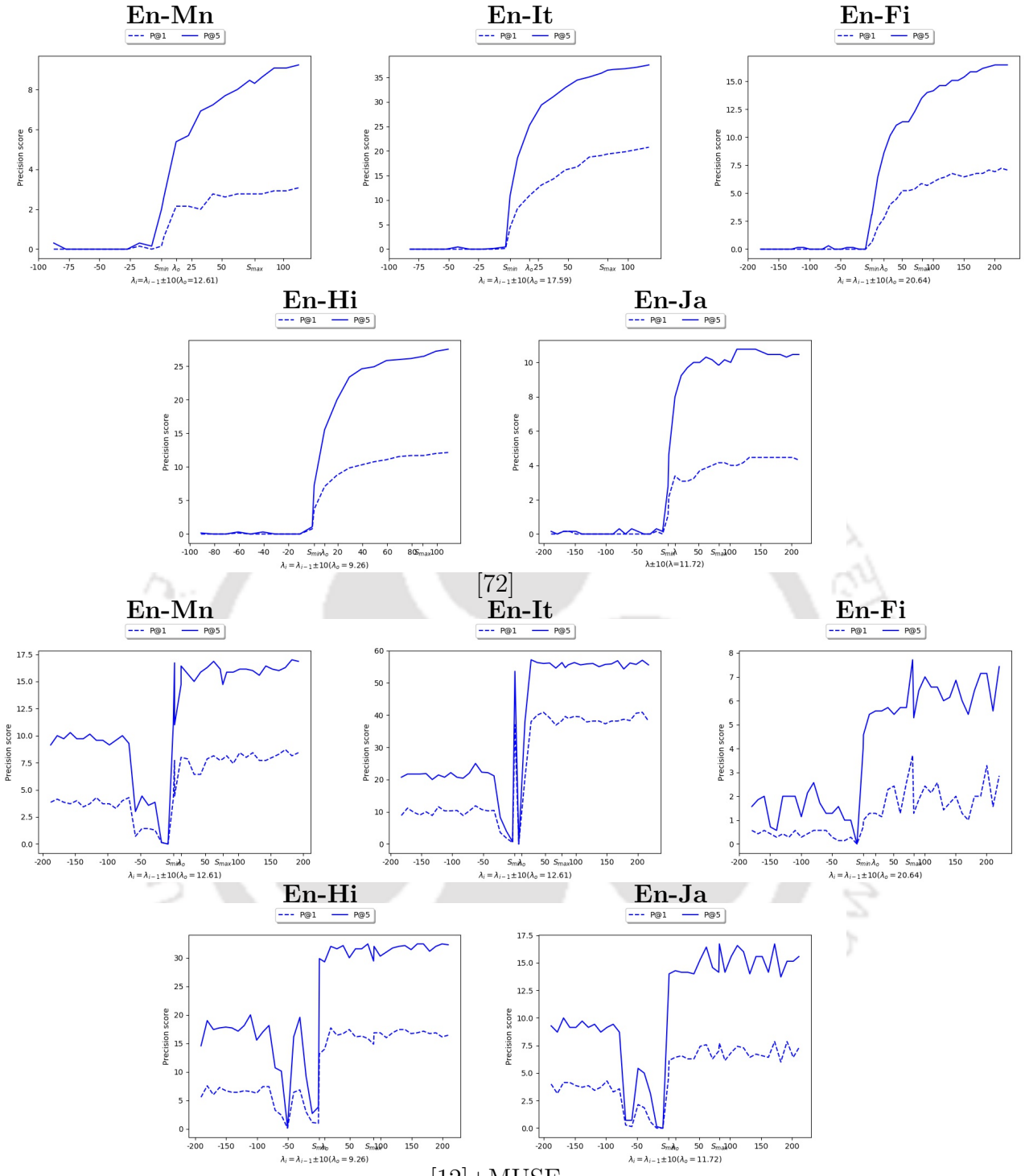
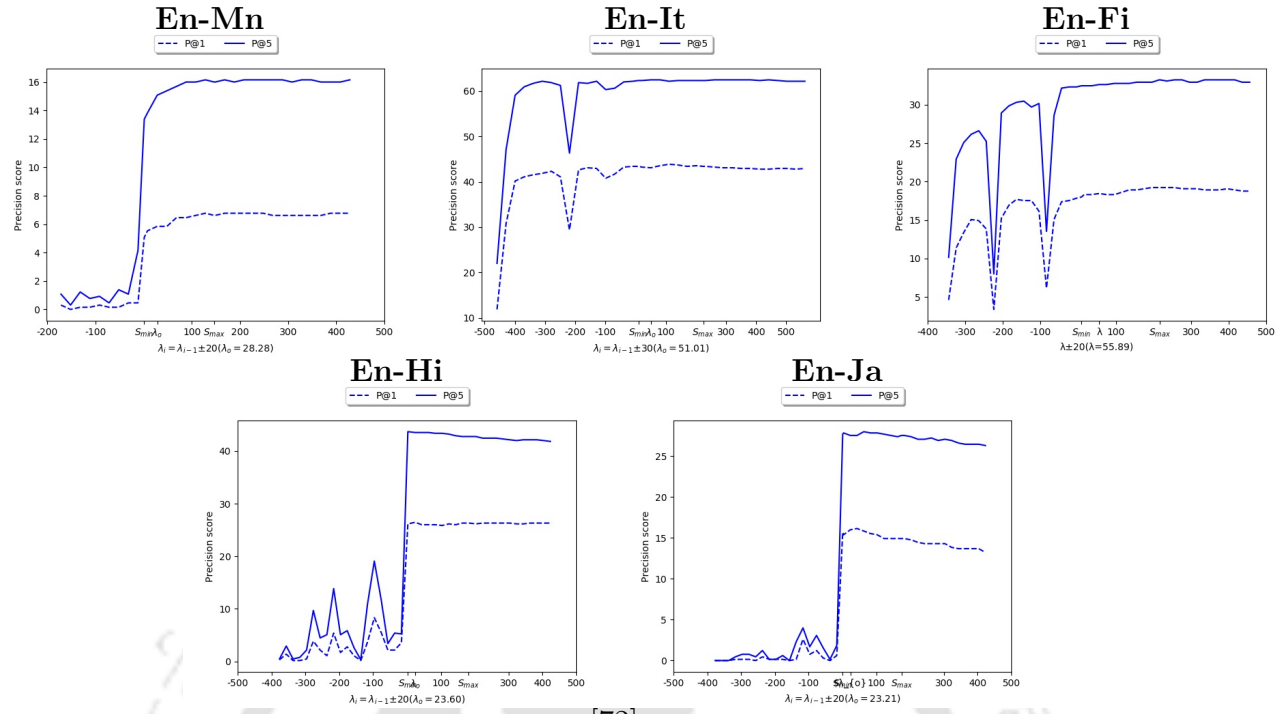


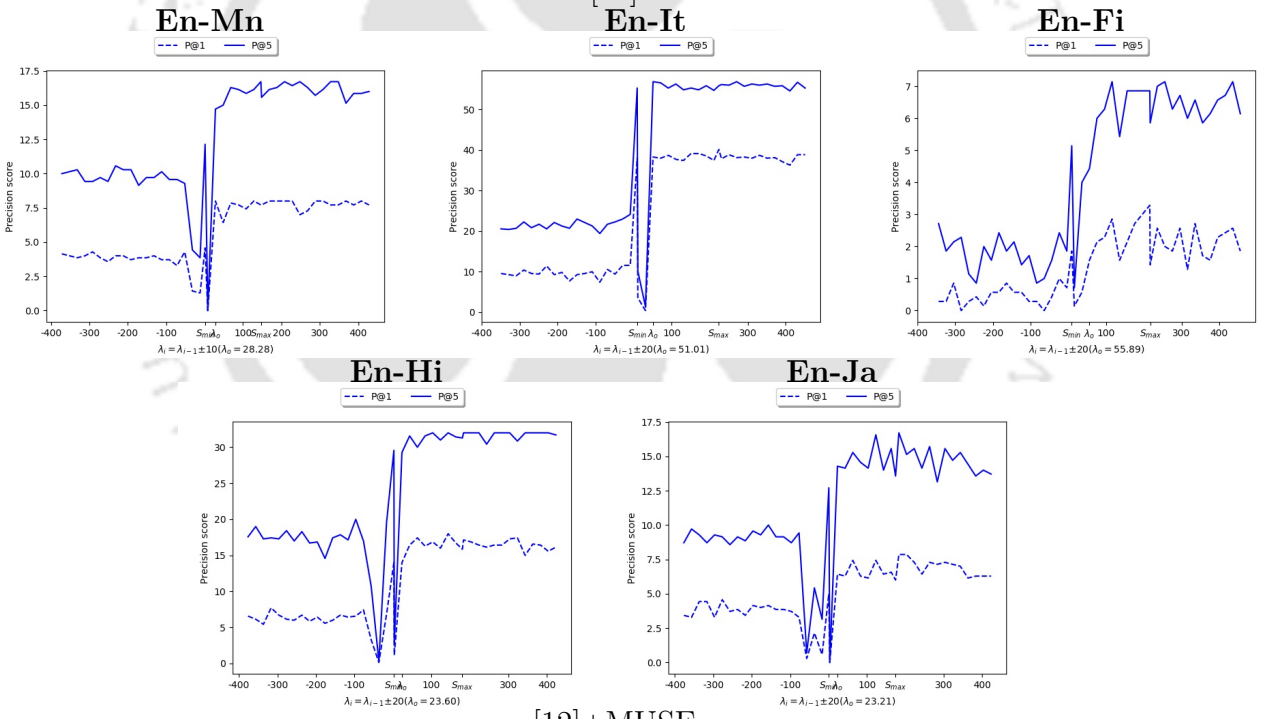
Figure 4.7: Evaluation on varying $\lambda_i = \lambda_{i-1} \pm \delta$ (10) using Degree Centrality at top-300

4.6.8 Isomorphic Similarity Test

To evaluate whether the proposed method enhances the isomorphism between the source and target embedding spaces, we compute the Eigenvector similarity for the embedding spaces generated by both the proposed and baseline methods. We then compare the Eigenvector



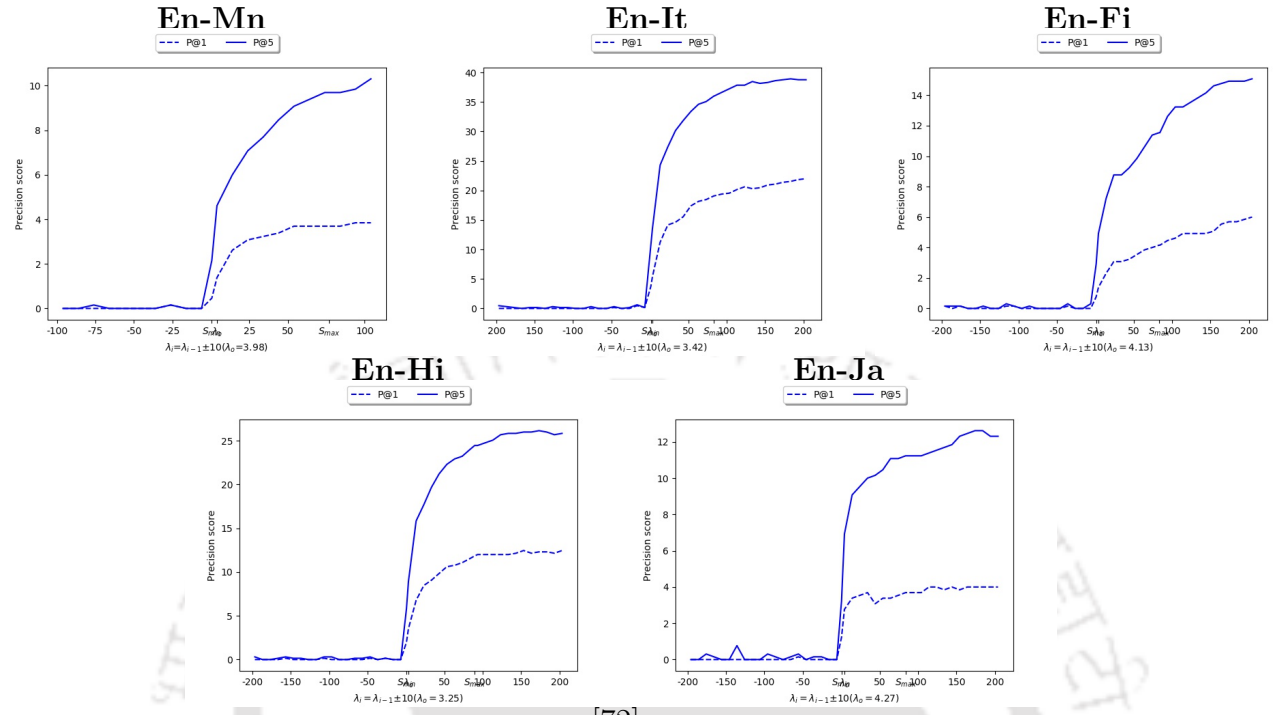
[72]



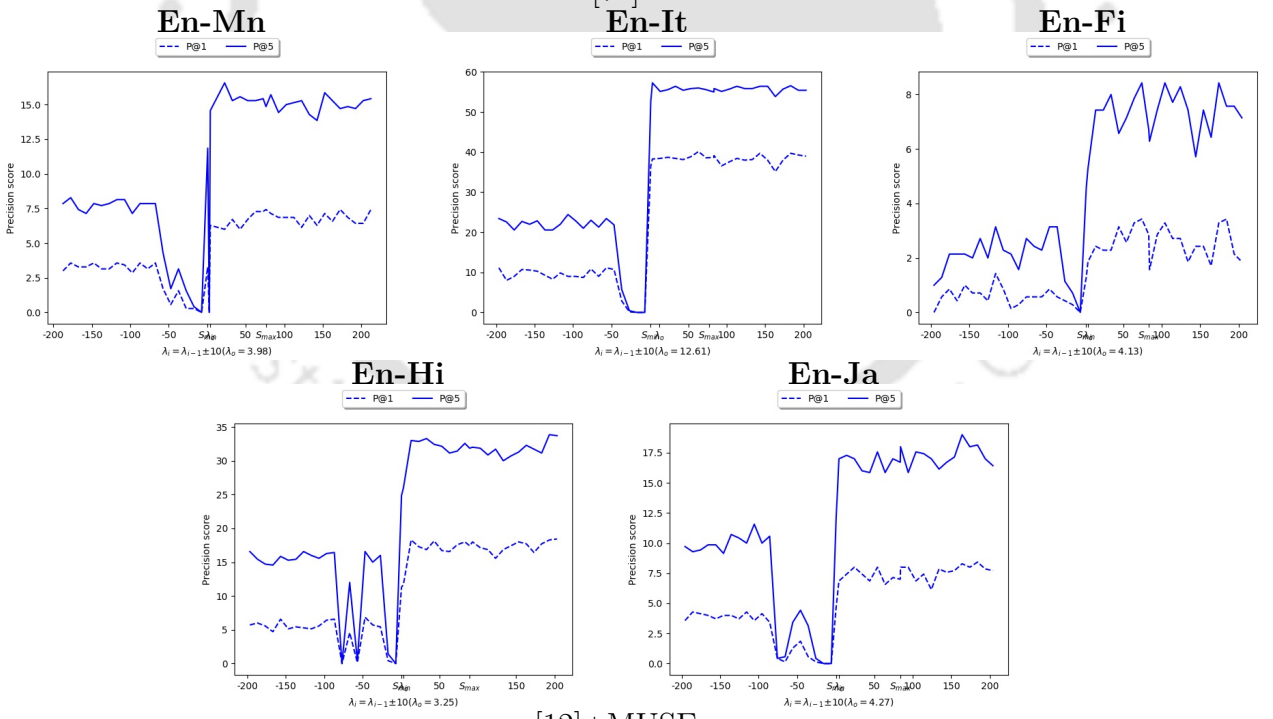
[12]+MUSE

Figure 4.8: Evaluation on varying $\lambda_i = \lambda_{i-1} \pm \delta$ (20) using Degree Centrality at top-4,200

similarity between these methods. Specifically, we select the second lowest Eigenvector for this comparison, as the first lowest Eigenvector does not convey meaningful information about the network properties. The second lowest Eigenvector encapsulates the connectivity and structural attributes of the graph and is particularly sensitive to subtle structural variations between graphs. The steps followed to determine Eigenvector similarity are as



[72]



[12]+MUSE

Figure 4.9: Evaluation on varying $\lambda_i = \lambda_{i-1} \pm \delta$ (10) using Eigenvector Centrality at top-300

Table 4.12: Comparison of Eigenvector similarity:Baseline Vs Proposed. Bold represent highest similarity score. FA: Frequency Align (Baseline), DCAR: Degree Centrality Align and Regularised, ECAR: Eigenvector Centrality Align and Regularised, LP: Language Pair

LP	V_λ	FA	DCAR	ECAR
En-Mn	2^{nd}	-0.004	-0.001	**0.002**
	3^{rd}	-0.010	-0.001	**0.001**
En-It	2^{nd}	-0.002	**0.003**	0.001
	3^{rd}	00.001	**0.002**	**0.002**
En-Fi	2^{nd}	-0.002	-0.001	**0.001**
	3^{rd}	-0.002	00.002	**0.004**
En-Hi	2^{nd}	-0.002	-0.001	**0.002**
	3^{rd}	-0.002	**0.002**	**0.002**
En-Ja	2^{nd}	00.006	0.0100	**0.020**
	3^{rd}	-0.003	0.0010	**0.002**

follows:

(i) Construct the similarity matrices for the source (S_s) and target (S_t) embedding spaces obtained from both the proposed and baseline methods.

(ii) Generate the corresponding Adjacency Matrices A_s and A_t from the similarity matrices S_s and S_t by applying a threshold of 0.2. An edge exists between two nodes in A if their cosine similarity score is at least 0.2.

(iii) Compute the Laplacian Matrices as $L_s = D_s - A_s$ and $L_t = D_t - A_t$, where D_s and D_t denote the diagonal degree matrices associated with A_s and A_t, respectively.

(iv) Determine the cosine similarity between the second and third lowest Eigenvectors of L_s and L_t.

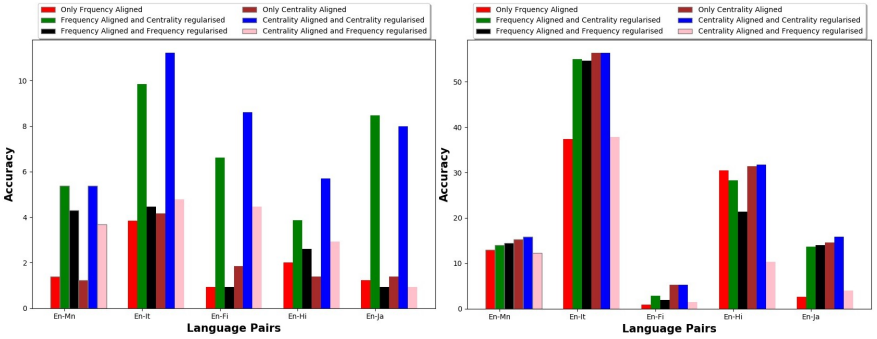
(v) Finally, compare the cosine similarity values obtained for the baseline and proposed methods.

Table 4.12 demonstrates that the proposed method achieves a higher Eigenvector similarity between the source and target embedding spaces. This suggests that the proposed method enhances the isomorphic properties of the source and target embedding spaces.

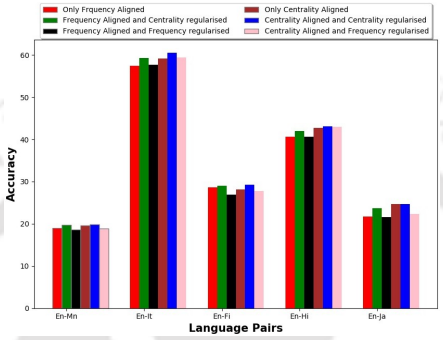
4.7 Ablation Test

To validate the significance of aligning dictionary pairs based on comparable centrality scores, six distinct ablation cases are analyzed using the methods proposed in [72], [12]+MUSE, and [12]+VECMAP.

The six ablation cases are as follows:



[72] [12]+MUSE



[12]+VECMAP

Figure 4.10: Ablation test using degree centrality in top-300 at P@5

(i) Alignment based solely on frequency. (**FA**)

(ii) Frequency-based alignment with additional centrality regularization. (**FACR**)

(iii) Frequency-based alignment with frequency regularization. (**FAFR**)

(iv) Alignment based solely on centrality. (**CA**)

(v) Centrality-based alignment with additional centrality regularization. (**CACR**)

(vi) Centrality-based alignment with frequency regularization. (**CAFR**)

From the experimental results presented in Figure 4.10 and Figure 4.11, it is evident that **CACR** achieved superior performance compared to the other five cases in 76.67

Instances where **CACR** performed equivalently or worse than other methods are outlined below:

(i) In En-Mn, **FACR** and **CACR** exhibited identical performance at top-300 in P@5 as reported in [72].

(ii) In En-Ja, **FACR** outperformed **CACR** at top-300 in P@5 according to [72].

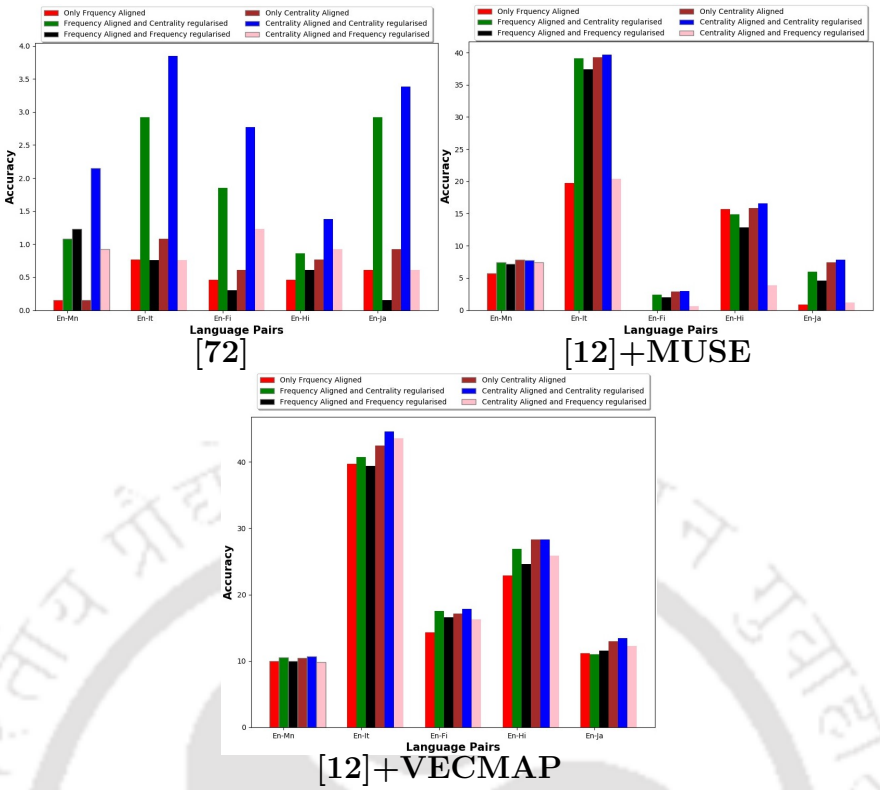


Figure 4.11: Ablation test using degree centrality (top-300 at P@1)

(iii) In En-Fi, **CA** and **CACR** achieved the same performance at top-300 in P@5 in [12]+MUSE.

(iv) In En-Ja, **CA** and **CACR** yielded identical results at top-300 in P@5 in [12]+VECMAP.

The ablation test results indicate that aligning dictionary pairs based on centrality scores and penalizing those with the least comparable centrality scores yields optimal performance.

Furthermore, the remaining ablation test results for the six cases specified in Section 4.7—[72] (top-300 at P@1, top-4,200 at P@1, and top-4,200 at P@5), [12]+MUSE (top-300 at P@1, top-4,200 at P@1, and top-4,200 at P@5), and [12]+VECMAP (top-300 at P@1, top-4,200 at P@1, and top-4,200 at P@5)—are illustrated in Figures 4.11, 4.12, and 4.13.

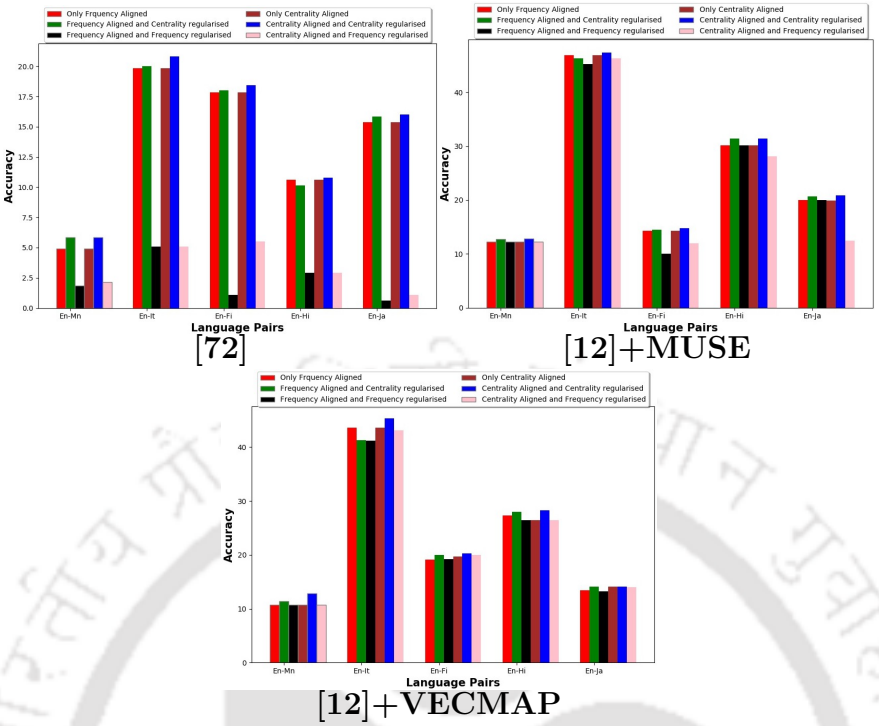


Figure 4.12: Ablation test using degree centrality (top-4,200 at P@1)

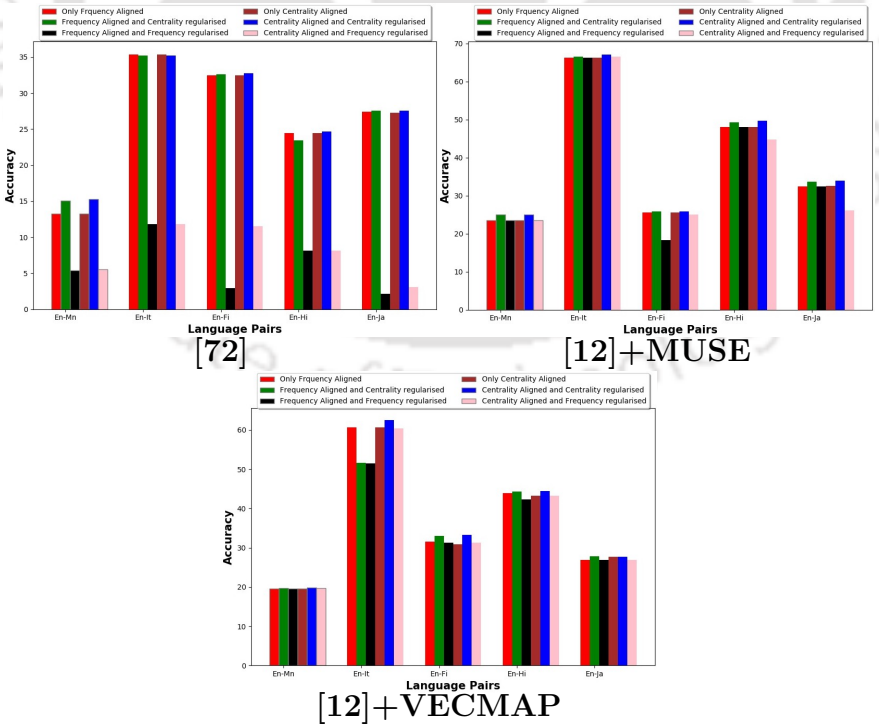


Figure 4.13: Ablation test using degree centrality (top-4,200 at P@5)

A few instances where **CACR** demonstrates equal or lower performance compared to other cases are outlined below:

(i) For En-Mn, **FACR** and **CACR** achieve identical performance in the top-4,200 at P@1 in [72].

(ii) In En-Mn, **CA** surpasses **CACR** in top-300 at P@1 in [12]+MUSE.

(iii) For En-Hi, **FACR** and **CACR** exhibit the same performance in the top-4,200 at P@1 in [12]+MUSE.

(iv) In En-Ja, **FACR** attains a slightly higher performance than **CACR** in the top-4,200 at P@5 for [12]+VECMAP.

Furthermore, Figures 4.11, 4.12, and 4.13 explicitly illustrate that aligning dictionary pairs based on centrality scores while penalizing those with the lowest comparable centrality scores leads to optimal performance.

4.8 Frequency Vs Centrality

A comparison between Frequency Aligned Frequency Regularised (**FAFR**) and Centrality Aligned Centrality Regularised (**CACR**) is presented by evaluating their performance on [72] and [12]+MUSE, ranging from top-300 to top-4,200. As illustrated in Figure 4.14, **CACR** consistently outperforms **FAFR** in most cases. This indicates that aligning dictionary pairs based on centrality while penalizing pairs with the lowest comparable centrality scores leads to superior performance compared to frequency-based alignment.

4.9 Error Analysis

The results of the regularized iterative normalization method [12] combined with MUSE, utilizing degree centrality, and the unregularized variant of the same method over the BDI task have been selected for error analysis. The translation correctness is determined by considering the five nearest neighbors. Table 4.13 presents the English words, their five closest translations along with their English meanings, and the corresponding ground truth Manipuri translations. For the English word **11**, the baseline method fails to retrieve the correct Manipuri translation within the five nearest neighbors. However, the proposed method successfully includes the correct translation ꯱꯱ in the nearest neighbors. Similarly, for the

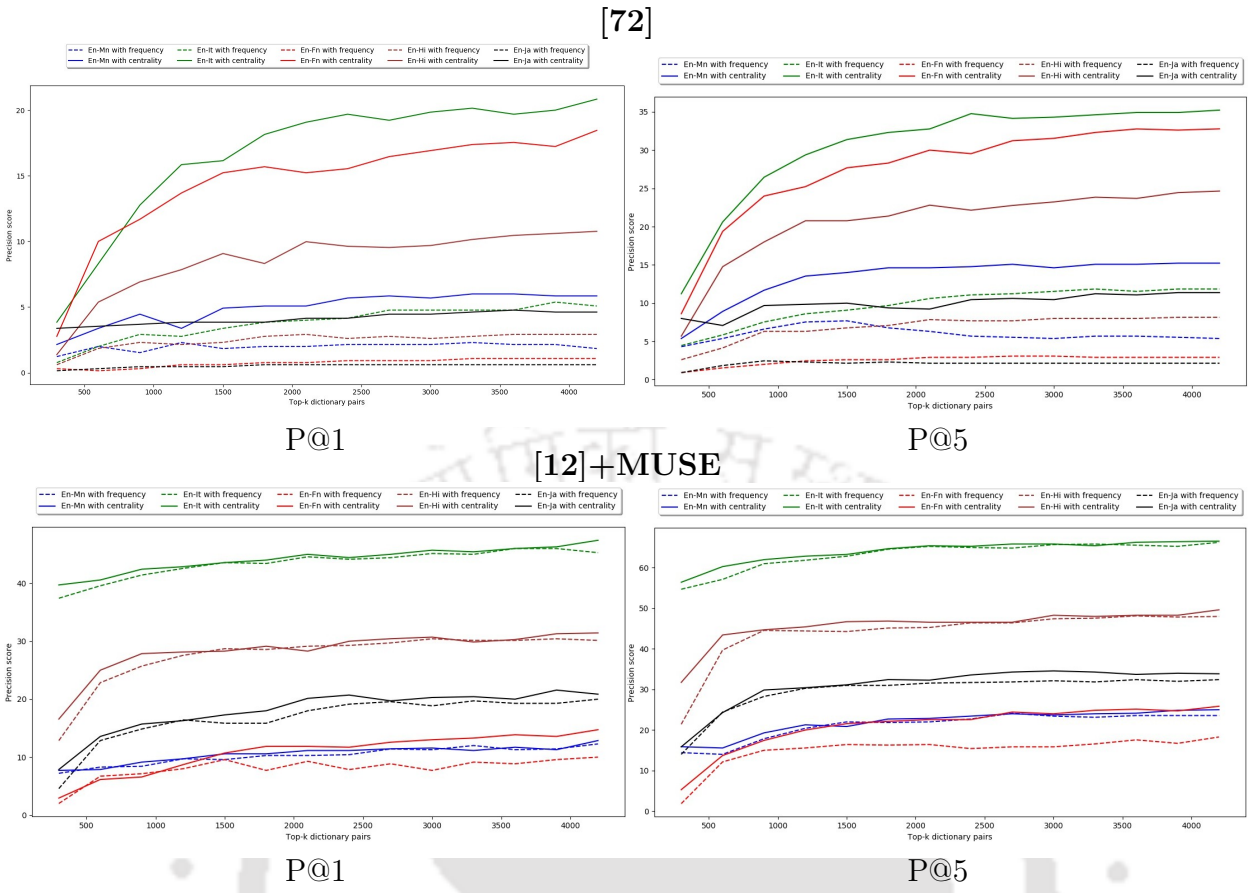


Figure 4.14: Frequency Vs Centrality evaluated in using degree centrality

English word **ball**, the correct Manipuri translation **বোল** appears only in the nearest neighbors generated by the proposed method. Additionally, the proposed method accurately translates **lighting** and **cinema** to **লাইতিং** and **সিনেমা**, respectively, whereas the baseline method fails to do so. In cases where exact matches to the ground truth are not found, the proposed method still provides semantically correct translations. For instance, the English word **establish** is transliterated as **ইস্তাব্লিশ** in the ground truth, but the proposed method instead generates **লিংখৎপা**, which conveys the correct meaning, unlike the baseline method. Similarly, for **shiv**, the proposed method predicts **সিব**, which, although not an exact match to the ground truth **শিব**, is a valid translation, while the baseline method fails to provide a correct alternative. However, both methods fail to correctly translate certain English words such as **name**, **outside**, **watching**, and **murderer**. In the case of **name**, the predicted translations **মমিংদা** (in his name) and **মিংদা** (in the name) include the suffix **দা**, leading to a slight semantic shift. Similarly, for **outside**, the predicted words **মপান্দগী** and **মপান্দা** contain the suffixes **দগী** and **দা**, altering the meaning from "from outside" to "at outside". Furthermore, the predicted words for **watching** and **murderer**, **য়েংবীয়ু** (have a look) and **মীহাৎতা** (murdered), include suffixes **বীয়ু** and **তা**, leading to semantic deviations from the respective ground truth words **য়েংলগা** (watching) and **মীহাৎপা** (murderer). This indicates that

English word	Baseline predicted	Proposed Predicted	Ground truth
11*	১২, ১৪, ৯, ৮, ৪ 12, 14, 9, 8, 4	১২, ১৪, ৮, **১১**, ৯ 12, 14, 8, 11, 9	**১১**
ball*	মাঙজীল, হাফ, গোলনা, পেনাল্টি, হেত-ত্রিক forward, half, by goal, penalty, hat-trick	পেনাল্টি, গোলনা, **বোল**, কাঙশিনখিবা, মাঙজীল penalty, by goal, ball, kick, forward	**বোল**
lighting*	কোল্ড, সিবিল, প্লান্তেসন, লোক্তাক্কী, রিজার্বর cold, pipe, plantation, for loktak, reservoir	কোল্ড, পাইপ, প্লান্তেসন, লোক্তাক্কী, **লাইতিং** cold, pipe, plantation, for loktak, lighting	**লাইতিং**
cinema*	হিন্দী, ক্লাসিকেল, হিন্দি, ফিল্ম, ম্যুজিক hindi, classical, hindi, film, music	হিন্দী, ক্লাসিকেল, **সিনেমা**, ম্যুজিক, নত hindi, classical, cinema, music, nat	**সিনেমা**
SAARC*	ব্রিক্স, বিমস্তেক, বিজিনেস, অনিমক্কা, বাইলেতরেল BRICS, BIMSTEC, business, both, biletral	বিমস্তেক, ব্রিক্স, এসিয়াগী, **সার্ক**, লৈবাকশিংগী BIMSTEC, BRICS, for asia, SAARC, for countries	**সার্ক**
bungalow*	সেক্রেতরিএত্তা, বঙ্গলোদা, রাজ, ভবন, চেম্বরদা at secretariat, at bungalow, raj, bhaban, at chamber	বঙ্গলোদা, রাজ, ভবন, সেক্রেতরিএত্তা, **বঙ্গলো** at bungalow, raj, bhaban, at secretariat, bungalow	**বঙ্গলো**
establish*	দিবেলপ, রিফোর্ম, অপগ্রেদ, ইনোবেসন, একচেঞ্জ develop, reform, upgrade, innovation, exchange	দিবেলপ, অপগ্রেদ, বিল্দিং, **লিংখৎপা**, রিফোর্ম develop, upgrade, building, establish, reform	**ইস্তাব্লিশ**
western*	তাইনরিবা, খা-নোংপোক, ম্যান্মাগা, এসিয়াগী, সাউথ neighbouring, south-eastern, with Myanmar, of Asia, south	তাইনরিবা, খা-নোংপোক, এসিয়াগী, সমুদ্র, **রেস্তর্ন** neighbouring, south-eastern, with Myanmar, of Asia, western	**রেষ্টর্ন**
shiv*	সেনা, মোর্চা, ভেটরান, এদিতর্স, অখিল sena, morcha, veteran, editors, akhil	সেনা, মোর্চা, **সিব**, ভেটরান, সিবা sena, morcha, shiv, veteran, shiva	**শিব**
name+	মিংদা, মমিংদা, ওঙ্গন, ফোতো, একাউন্ত in the name, in his name, option, photo, account	ওঙ্গন, মমিংদা, মিংদা, ফোতো, লাইরিকশিং option, in the name, in his name, photo, books	**মমিং**
outside+	মপান্দগী, তারগসু, লমদগী, মপান্দা, পেসেঞ্জরশিং from outside, listening, that place, at outside, passengers	মপান্দগী, লমদগী, মপান্দা, তারগসু, পেসেঞ্জরশিং from outside, that place, at outside, listening, passengers	**মপান**
watching+	য়েংবীয়ু, ইবদা, খঙডি, নখোয়না, ঐবু have a look, in writing, I know, all of you, I	য়েংবীয়ু, ইবদা, খঙডি, ঐবু, নখোয়না have a look, in writing, I know, I, all of you	**য়েংলগা**
murderer+	ফিজম, ফরুক, ধনবির, মীহাৎতা, প্রবিস Phejam, Farukh, Dhanabir, in death, Pravish	ফিজম, ফরুক, ধনবির, প্রবিস, মীহাৎতা Phejam, Farukh, Dhanabir, Pravish, in death	**মীহাৎপা**
date+	মীরেপ, কেন্দিদেত, মেম্বর, মীরেপশিং, কেন্দিদেট candidate, candidate, member, candidates, candidate	মীরেপ, কেন্দিদেত, মেম্বর, মীরেপশিং, কেন্দিদেট candidate, candidate, member, candidates, candidate	**তারিখ**
interest+	নৈনবা, অরুবা, খন্ন, নখোয়না, তশেং discussion, critical, discuss, all of you, real	নৈনবা, অরুবা, খন্ন, নুংঙাইবা, তশেং discussion, critical, discuss, happy, real	**ইন্তরেস্ততা**

Table 4.13: Baseline Vs Proposed in terms of five nearest predicted translation of English word. The English meaning of the predicted words are given in the next row. *: shows strength of the proposed method and +: shows weakness of the proposed method. Bold shows correct translation

neither the baseline nor the proposed method effectively handles morphological inflection. Additionally, some predicted words are entirely unrelated to the ground truth translations. For instance, the English words **date** and **interest** are translated into Manipuri words that are semantically different from the expected ground truth translations.

4.10 Summary and Future work

Experimental observations indicate that diverse language pairs often yield poor dictionary pairs, which do not significantly contribute to cross-lingual alignment. Instead of selecting dictionary pairs solely based on the frequency of the source language, it is more effective to choose pairs with comparable centrality measures. Centrality measures provide valuable insights into word coverage, implicitly suggesting that the neighbors of a source word should exhibit semantic similarity to those of the target word. Therefore, down-weighting dictionary pairs with the least comparable centrality measures enhances performance in the Bilingual Dictionary Induction (BDI) task.

Furthermore, experiments conducted on Cross-lingual Sentence Retrieval Tasks and Machine Translation demonstrate that our proposed method surpasses the baseline approach. This study also establishes that regularization does not inherently increase hubness or degrade

BDI performance. Additionally, the regularization term λ is not arbitrary; penalizing dictionary pairs based on their centrality difference factor contributes to improved performance in BDI tasks. Word embeddings derived from multilingual LLMs, such as mBERT, do not exhibit lexical properties equivalent to those of static word embeddings like Word2Vec and FastText. Consequently, for resource-constrained languages like Manipuri, the BDI task plays a crucial role in developing other NLP tools. Thus, projection-based methods like VECMAP, which yield superior BDI performance using static embeddings, remain essential for low-resource languages.

Evaluation using an unsupervised approach, such as unsupervised VECMAP[42] with regularization, fails to enhance BDI task performance. Instead, it leads to performance degradation, likely due to the poor quality of the initial seed dictionary generated by the unsupervised method. In this study, only Degree Centrality and Eigenvector Centrality were considered, leaving the exploration of other potentially more suitable centrality measures as an open avenue for future research. Beyond dictionary selection, leveraging language-specific linguistic features, such as morphology and sentence structure, could further enhance performance in downstream NLP tasks. Additionally, evaluating other downstream NLP tasks would provide deeper insights into the quality of cross-lingual embeddings with regularization.

Chapter 5

A Novel Morphology Aware CLWE framework

Bilingual Dictionary Induction (BDI) task induced by cross-lingual word embedding usually performs poorly when the language pairs are structurally non-isomorphic, like English-Manipuri. The performance further degrades when one of the languages in the pair is morphologically rich. Earlier work shows that segmenting morpheme (root) and suffix/prefix using a morphological analyzer slightly improves cross-lingual word embedding in the BDI task. In this chapter, we proposed a novel contrastive learning method that exploits the rich morphological nature of a language to our advantage without segmenting root words and suffixes/prefixes. The proposed contrastive learning pulls the source and target words together in a bilingual dictionary and brings the target word with another target word that shares the same root. From various experimental observations over four language pairs, it is evident that the proposed method outperforms baseline methods in the BDI, Machine Translation (MT), and Cross-lingual Sentence Retrieval Task (CSRT).

5.1 Introduction

Among the popular cross-lingual word embedding approaches, the study reported in [69, 85] observes that the orthogonal cross-lingual mapping method, such as VecMap[23], still outperforms the multilingual language model (LM), such as mBERT [55], in the bilingual dictionary induction task (BDI). Though VecMap and its variant [42] are found to be performing well for isomorphic language pairs, they are not that effective for distant language pairs. The performance worsens when one language is morphologically rich, like English-Finnish and English-Turkish [41].

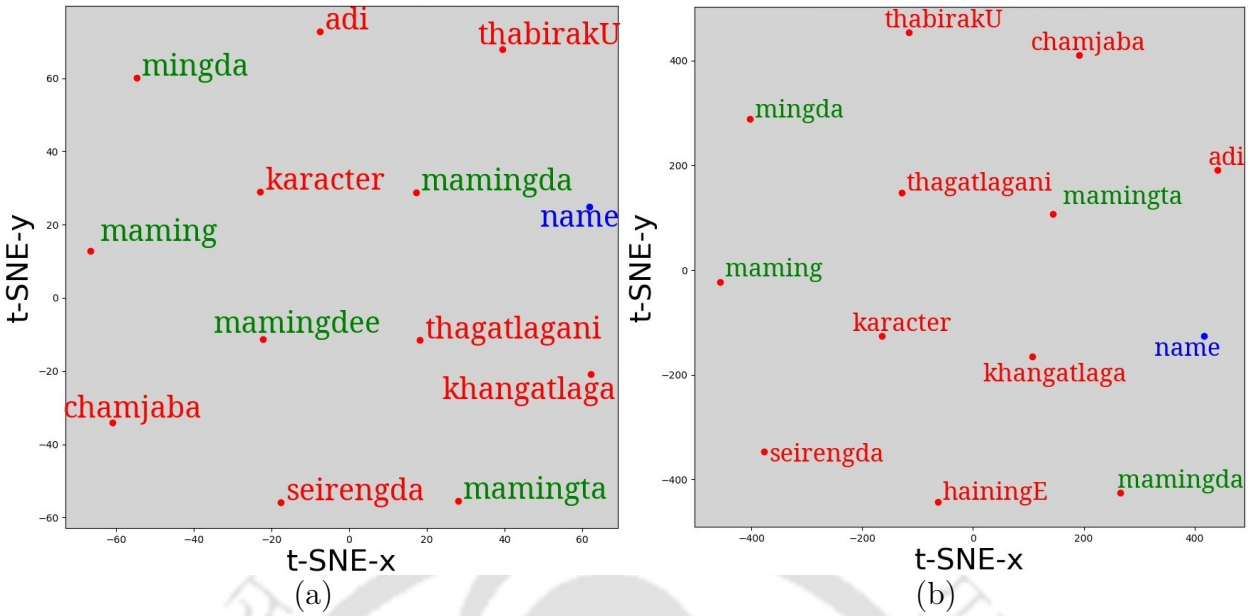


Figure 5.1: A t-SNE visualisation in En-Mn. Same root Manipuri words are plotted in green, English word is plotted in blue. Semantically unrelated word closed to Manipuri word "maming" are plotted in red. (a)VecMap (b)Li et al.[1]

Morphological-rich language causes inaccurate mappings while handling complex words because the existing CLWE models do not exploit sub-word level morphological information while aligning complex words [27]. It is observed that segmenting morpheme (root) and suffixes help in improving BDI performance [27, 28]. However, effective stemmers are only available for some languages, especially rich-resource ones. Therefore, we need CLWE models to enhance BDI performance for morphological-rich languages without root and suffix segmentation. Semantically similar words can be formed from a given root word in a morphologically rich language. For an effective cross-lingual embedding, such words with the same root must be brought closer to each other. A study in [86] shows an increase in the monolingual downstream tasks by bringing words with the same root closer. However, so far, CLWE embedding methods such as VecMap[23] and [1] fail to handle the inaccurate mapping due to complex morphological properties in the BDI task. An illustrative example is given in Figure 5.1 (a). For an English word **name**, the top five BDI translations in VecMap are **mamingta** (in the name), **mamingda** (at his name), **thagatlagani** (will be elected), **khangatlaga** (selected one), **karacter** (character). However, the correct translation is **maming** (name). The unrelated words **thagatlagani**, **karacter**, and **khangatlaga** are pulled among the top 5 nearest neighbors. A similar trend is also observed in the CLWE proposed in [1] as shown in Figure 5.1 (b). This CLWE characteristic hinders the performance of BDI. To address the above issue, words with the same root are pulled closer, pushing away unrelated words from 5-nearest neighbors.

Motivated by the above concerns, this chapter proposes MACE (Morphology Aware Cross-lingual Embedding using Contrastive Learning), which brings target language words with the same root closer and pushes target words with different roots apart using *contrastive learning*. A study in [1] has also observed an improved performance while applying contrastive learning for BDI tasks. However, the authors in [1] apply contrastive learning only at the positive word pairs (dictionary words) and negative word pairs (non-dictionary pair words). Unlike [1], we proposed additional contrastive learning among target words with the same root as positive pairs and those with different roots as negative pairs. Like [1], we used VecMap and mBERT/IndicBert[55, 57] as the orthogonal mapping and pre-trained language model counterparts for applying contrastive learning. For the first time, the large language model (LLM), mT5[87], is trained in the continued training process [88] with Manipuri data and is later used in the proposed method setting for English-Manipuri BDI evaluation. From various experimental observations, it is evident that the proposed model outperforms all its counterparts in the downstream applications - BDI, Cross-lingual Sentence Retrieval Task, and Machine Translation.

5.1.1 Contribution

The major contributions of this chapter are:

- Proposed CLWE contrastive learning to bring target words with the same root closer and target words with different roots apart in a CLWE framework.
- The proposed framework has been evaluated with English as the source language and four morphologically rich languages as the target language and observes superior performances over the baseline method.

5.2 Related Studies

Initial mapping-based CLWE model in [72] developed a regression model to induce a linear mapping function using a bilingual seed dictionary. The method proposed in [3] and its variant [42] introduce a closed-form solution popularly known as VECMAP.[1] Method proposed in [89] that uses centrality-aligned ridge regression-based orthogonal mapping fails to handle morphological complex words in BDI task. An empirical study in [76] shows that both supervised and unsupervised approaches perform poorly in morphologically rich

[1] https://github.com/artetxem/vecmap

languages like Finnish and Turkish in the BDI task. A morpheme-based approach that Segments morpheme (root) and suffixes slightly improve the BDI performance [27, 28]. A recent study [47] introduced 40 morphologically complete dictionaries and revealed that the performance in BDI degrades severely in the case of less frequent inflected words. Another study [90] proposed a morphologically aware probability model for bilingual lexicon induction, which jointly models the lexeme translation and inflectional morphology. The method proposed in [86] shows an increase in monolingual downstream tasks by bringing words with the same root closer. With the success of contrastive learning, contrastive learning-based cross-lingual word representation [1] and cross-lingual sentence embedding [91] are proposed. Still, the method proposed in [1] doesn't perform well for morphologically rich languages like Finnish and Turkish. Few-short prompting method using llama proposed in [18] shows state-of-the-art BDI scores in many language pairs, excluding pairs when one language is morphologically rich.

5.3 Proposed Method

The proposed method combines two cross-lingual embedding models within a contrastive learning framework. It consists of three stages, as illustrated in Figure 5.2 - (i) *Generation of positive and negative pairs for contrastive learning*, (ii) *Fine-tuning both VecMap and LM/LLM WEs with contrastive loss using the positive and negative pairs to bring words with the same root closer*, and (iii) *ensembling of the fine-tuned representations to take advantage of both the spectral representation of corpus and linguistic characteristics of languages*. We select state-of-the-art VecMap for orthogonal projection and pre-trained language model/LLM (IndicBERT/mBERT/mT5). Both embeddings of VecMap and pre-trained language model/LLM are fine-tuned with an additional contrastive learning process using positive (words with the same root) and negative (words with different root) samples generated using a bilingual dictionary and monolingual embeddings. Positive and negative sample generation details are reported in section 5.3.1. Like in the study [1], the outcomes of the two contrastively fine-tuned embeddings are ensembled by learning another linear orthogonal matrix $\hat{W} \epsilon \mathbb{R}^{300 \times 768}$ as proposed in [92, 93] (details in section 5.3.3). Thus, the proposed contrastive learning method exploits the orthogonal projection-based CLWEs and multilingual LM/LLM, brings words with the same root closer, and pushes words with different roots apart. The details of the components of the proposed framework are discussed below.

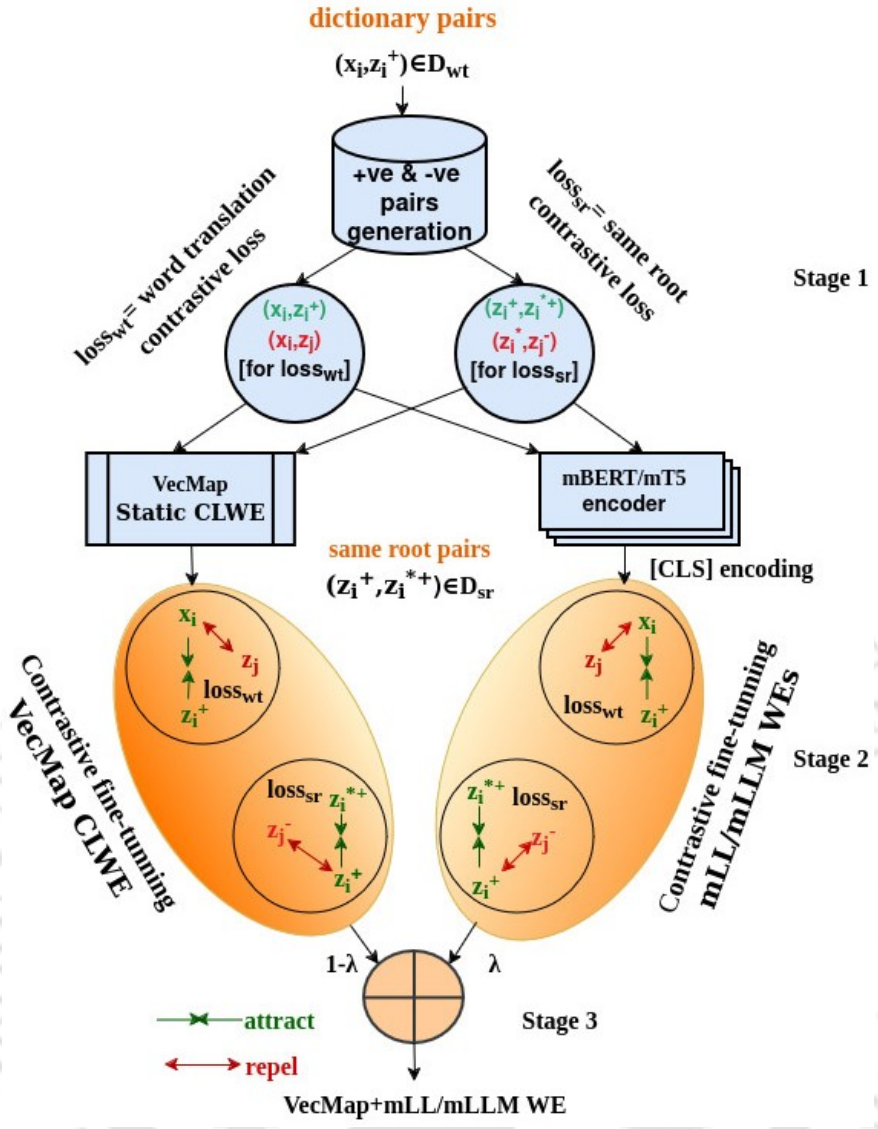


Figure 5.2: Proposed Method Architecture

5.3.1 Generating Positive and Negative pairs

As mentioned above, the proposed model applies contrastive loss over the CLWEs. We generate positive and negative word pairs to use in contrastive loss. For generating positive and negative word pairs, we first consider a seed bilingual dictionary for generating positive and negative sample pairs (the same seed dictionary used in Vecmap). If $z = f(x)$ defines a dictionary word of x, a bilingual seed dictionary D_{wt} can be defined as $D_{wt} = \{(x, z)|x \in X, z \in Z, z = f(x)\}$ where X and Z are the words in source and target languages, respectively. We generate positive and negative pairs for two contrastive losses as given below.

(i) Word Translation Contrastive loss, $loss_{wt}$

(ii) Same Root Contrastive loss, $loss_{sr}$

5.3.1.1 Loss_{wt}

Like in the contrastive learning approach [1], for a given positive pair $(x_i, z_i^+)\epsilon D_{wt}$, a hard negative set $S_z^- = \{z_j \mid (x_i, z_i^+)\epsilon D_{wt}, z_j \neq z_i^+, z_j \notin NN(x_i)\}$ is generated where $NN(x_i)$ is the nearest neighbors of x_i from VecMap embedding of target language excluding z_i^+. We generate the negative pair set $S_x^- = \{x_j \mid (x_i, z_i^+)\epsilon D_{wt}, x_j \neq x_i, x_j \notin NN(z_i^+)\}$ For target-to-source translation, where $NN(z_i^+)$ is the nearest neighbors of z_i^+ from VecMap embedding of source language excluding x_i.

5.3.1.2 Loss_{sr}

For each word pair $(x_i, z_i^+) \in D_{wt}$, we further generate a positive set of z_i^+ (words in target language having same root with z_i^+). Let D_{sr} be the set of such positive pairs. We have used edit distance to generate the same root target word pairs. For $(x_i, z_i^+) \in D_{wt}$, we now define a negative set of z_i^+ (words in the target language having different root with z_i^+) using the D_{sr} defined above. If $(z_i^+, z_i^{*+}) \in D_{sr}$, the hard negative set of z_i^+ is defined as below:
$S_{sr}^- = \{z_j^- \mid (z_i^+, z_j^-) \notin D_{sr}, z_j^- \in Z, z_j^- \notin NN(z_i^{*+})\}$
where $NN(z_i^{*+})$ are the nearest neighbours of z_i^{*+} as described in the above section 5.3.1.1. In the above formulation of D_{sr}, for a given target word, it considers all the words with the same root as the positive set of the target word. We refer to it as multiple +ve. The larger the number of samples in contrastive loss, the higher the computational cost. To investigate the effectiveness of the model with a smaller number of the positive sets, we further formulate single +ve where, for one target word, it considers only one random +ve target word that has the same root as shown in Figure 5.3.

5.3.2 Cross-lingual Contrastive learning

Once we obtain the set of positive and negative samples for the word pairs in the seed dictionary, we apply the following contrastive fine-tuning over VecMap and pre-trained LM/LLM.

(i) Fine-tuning static VecMap WEs

(ii) Fine-tuning pre-trained multilingual LM/LLM

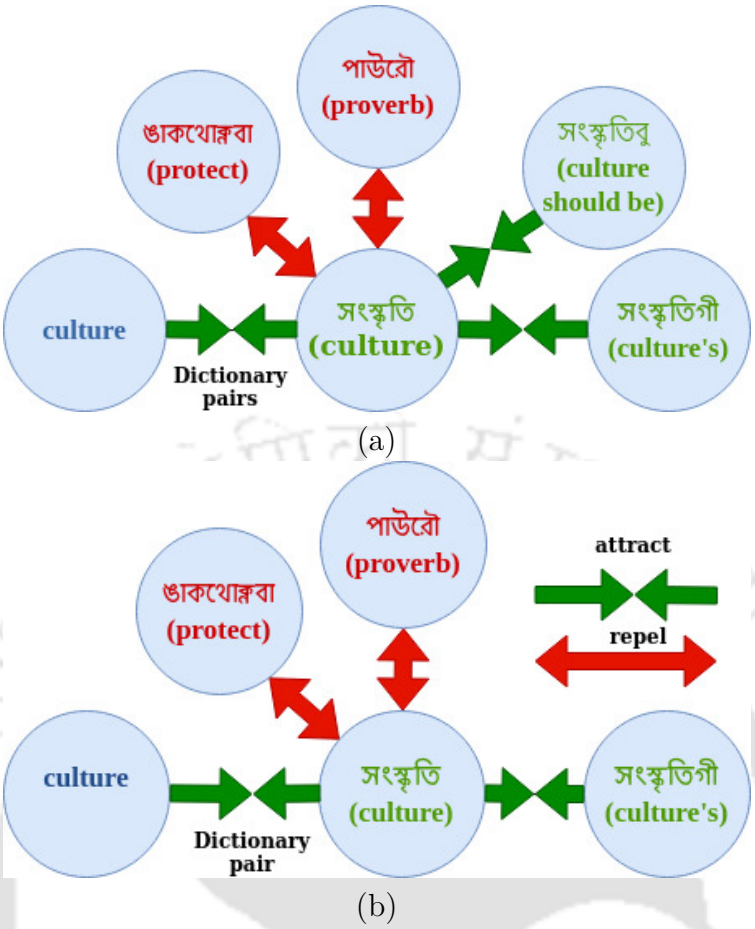


Figure 5.3: Multi +ve and Single +ve, words with same root word (+ve pair) are given in green color, -ve pairs are given in red color and English word is given in blue color. (a)multi +ve. (b)single +ve

5.3.2.1 Fine-tuning static VecMap WEs

We use the contrastive learning objective as described in [1] that uses two linear transformation matrices W_x and W_z obtained from VecMap[42], where W_x is the transformation matrix for source language and W_z is the transformation matrix for target language. The contrastive learning objective function is given below:

$$loss_{wt} = \frac{sim(x_i, z_i^+)}{\sum\limits_{z_j \epsilon \left\{z_i^+\right\} \cup S_z^-} sim(x_i, z_j) + \sum\limits_{x_j \epsilon S_x^-} sim(x_j, z_i^+)} \tag{5.1}$$

$$sim(x_i, z_j) = \exp^{\cos(x_i W_x, z_j W_z)/\tau} \tag{5.2}$$

$$\min_{W_x,W_z} -\mathbb{E}_{(x_i,z_i^+)\in D_{wt}} \log(\text{loss}_{wt}) \tag{5.3}$$

where τ is the standard temperature parameter.

In addition to $loss_{wt}$, our proposed method introduces another contrastive learning objective $loss_{sr}$, which brings target language words with the same root closer and target language words with different roots apart. The above-described contrastive learning objective function is given below:

$$loss_{sr} = \frac{sim_{sr}(z_i^+, z_i^{*+})}{\sum\limits_{z_j \epsilon \left\{z_i^{*+}\right\} \cup S_{sr}^-} sim_{sr}(z_i^+, z_j)} \tag{5.4}$$

$$sim_{sr}(z_i^+, z_j) = \exp^{\cos(z_i^+ W_z, z_j W_z)/\tau} \tag{5.5}$$

The combined contrastive learning objective function for single +ve and multi +ve scenarios are given in equations 5.6 and 5.7 respectively. Here β_1 is a hyperparameter. M is a set of target language words with the same root as z_i^+.

$$\min_{W_x,W_z} -\Big[\mathbb{E}_{(x_i,z_i^+)\in D_{wt}} \log(\text{loss}_{wt}) + \beta_1 \mathbb{E}_{(z_i^+,z_i^{*+})\in D_{sr}} \log(\text{loss}_{sr})\Big] \tag{5.6}$$

$$\min_{W_x,W_z} -\Big[\mathbb{E}_{(x_i,z_i^+)\in D_{wt}} \log(\text{loss}_{wt}) + \beta_1 \mathbb{E}_{(z_i^+,z_i^{*+})\in D_{sr}} \sum_{z_i^{*+}\epsilon M} \log(\text{loss}_{sr})\Big] \tag{5.7}$$

5.3.2.2 Fine-tuning pretrained multilingual LM/LLM

We use pre-trained $mBERT_{base}$, IndicBERT (for En-Mn), and mT5 model for contrastive fine-tuning. Given a pair $(x_i, z_i^+)\epsilon D_{wt}$, x_i and z_i^+ are tokenized using mBERT tokenizer, giving the following sub-words sequences: $[CLS]s_1x_i....s_nx_i[SEP]$, $[CLS]s_1z_i^+....s_nz_i^+[SEP]$, $n \geq 1$. $[CLS]$ and $[SEP]$ are special tokens as reported in [69]. The mBERT encoding function f_θ takes the sequence as input and gives the representation of the $[CLS]$ token in the last transformer layer as the representation of x_i and z_i^+ respectively. We also extract the representation of z_i^{*+}, where $(z_i^+, z_i^{*+})\epsilon D_{sr}$ in the similarly way. The same contrastive

InfoNCE $loss_{wt}$ and $loss_{sr}$ given in equations 5.1, 5.4 respectively are used.

$$sim(x_i, z_j) = \exp^{\cos(f_\theta(x_i),\, f_\theta(z_j))/\tau} \tag{5.8}$$

$$sim_{sr}(z_i^+, z_j) = \exp^{\cos(f_\theta(z_i^+),\, f_\theta(z_j))/\tau} \tag{5.9}$$

The final contrastive learning objective function that fine-tune LM/LLM parameters θ for single +ve and multi +ve scenarios are given in equations 5.10 and 5.11 respectively. Here β_2 is a hyperparameter.

$$\min_\theta - \Big[\mathbb{E}_{(x_i, z_i^+)\in D_{wt}} \log(\text{loss}_{wt}) + \beta_2 \mathbb{E}_{(z_i^+, z_i^{*+})\in D_{sr}} \log(\text{loss}_{sr})\Big] \tag{5.10}$$

$$\min_\theta - \Big[\mathbb{E}_{(x_i, z_i^+)\in D_{wt}} \log(\text{loss}_{wt}) + \beta_2 \mathbb{E}_{(z_i^+, z_i^{*+})\in D_{sr}} \sum_{z_i^{*+}\epsilon M} \log(\text{loss}_{sr})\Big] \tag{5.11}$$

5.3.3 Fusing Static WE and LM/LLM

The contrastive fine-tuned static VecMap WEs are fused with contrastive fine-tuned LM/LLM WEs to leverage the quality of cross-lingual embeddings. The fusing is achieved by learning a linear orthogonal matrix $\hat{W}\epsilon\mathbb{R}^{300\times768}$ that maps 300-dimension contrastively fine-tuned static cross-lingual WE space into 768-dimension (512 for mT5-small) cross-lingual LM/LLM WE space[92, 93]. The final fused representation of an input word w is a linear combination given below:

$$(1-\lambda)\frac{v_w\hat{W}}{\left\|v_w\hat{W}\right\|_2} + \lambda\frac{f_\theta(w)}{\|f_\theta(w)\|_2} \tag{5.12}$$

where v_w is the contrastively fine-tuned static WE and λ interpolation hyper-parameter.

5.4 Dataset

This study considers four language pairs - English-Finnish (En-Fi), English-Turkish (En-Tr), English-Tamil (En-Ta), and English-Manipuri (En-Mn). For En-Fi, the Europarl[2] parallel corpus [54] extracted from the proceedings of the European Parliament is used. Parallel corpus provided in MaCoCu-tr-en 2.0[94] is used for En-Tr. En-Ta parallel corpus from Bharat Parallel Corpus Collection (BPCC), AI4BHARAT[3] is used. Manipuri, a low-resource language, does not have a sufficient English-Manipuri (En-Mn) parallel corpus. A document-level corpus comparable is used for En-Mn. The comparable corpus is extracted from two prominent Manipuri online news article platforms: Sangai Express[4] and Poknafam[5]. English-Manipuri comparable corpus is also extracted from news updates posted on PMIndia[6] [79]. The details of the data is shown in Table 5.1. Using a seed dictionary, CLWE methods generally project monolingual source and target language word embeddings to a shared space. This study considers the bilingual dictionary available at Directorate of Language Planning and Implementation, Government of Manipur [7] for En-Mn language pair, and the MUSE[8] library for the En-Tr, En-Fi, and En-Ta language pairs.

[2] https://www.statmt.org/europarl/
[3] https://ai4bharat.iitm.ac.in/bpcc/
[4] https://www.thesangaiexpress.com/index.html
[5] http://www.poknapham.in
[6] https://www.pmindia.gov.in/en/
[7] https://www.dlpi.mn.gov.in/en/
[8] https://github.com/facebookresearch/MUSE

Table 5.1: The statistics of data, **LP**: Language Pairs

LP	Platform	sentences		words		unique words	
En-Mn	**Sangai Express**	**En**	**Mn**	**En**	**Mn**	**En**	**Mn**
	+Poknafam+PMI	645208	458436	14.1M	19.3M	160642	43599
En-Fi	**European Parliament**	**En**	**Fi**	**En**	**Fi**	**En**	**Fi**
		1.92 M	1.92M	47.4M	32.2M	151017	219976
En-Ta	**AI for Bharat**	**En**	**Ta**	**En**	**Ta**	**En**	**Ta**
		442776	442776	10.3M	8.4M	79518	314452
En-Tr	**MaCoCu-tr-en 2.0**	**En**	**Tr**	**En**	**Tr**	**En**	**Tr**
		1.6 M	1.6M	55.0M	51.5M	411397	884161

5.5 Experimental Setup

The data mentioned in section 5.4 are used to generate respective language monolingual word embeddings. We generate monolingual word embedding using fastText[95] with dimension=300, window size=5, min_{count}=5, min_{n-gram}=3 and max_{n-gram}=6. The VecMap embeddings are generated using the respective monolingual fastText embeddings. For contrastive fine-tuning of LM/LLM, we used pre-trained $mBERT_{base}$ (En-Fi, En-Tr, and En-Ta)[55], IndicBERTv2-MLM-only (En-Mn)[57], and $mT5_{small}$[87] (continue training with Manipuri data) with the embedding dimension size 512. A training dictionary of 3500 pairs and 700 testing pairs are generated for all language pairs. We create D_{sr} as described in section 5.3.1.2 with edit distance=1, 2, 3. Each experiment with the same parameter is run five times, and the mean and standardized deviation are reported in the experimental results. We consider only the morphologically rich language given in section 5.4 as the target language. Non-morphological rich language will give poor semantically closed same-root word pair with edit=1.

5.5.1 Contrastive learning parameter

The hyperparameter values are N_{iter}=2, N_{wt}=50, N_{sr}=50, N_{iter} is the no of iterations in VecMap. N_{wt} is the no of negative samples for a positive pair in $loss_{wt}$. N_{sr} is the no of negative samples for a positive pair in $loss_{sr}$. For fine-tuning static WEs, we used an SGD optimizer with a learning rate 1.5. AdamW[56] optimizer with a learning rate of 2e-5 is used for fine-tuning LM/LLM WEs. Both static and LM/LLM WEs are fine-tuned for five epochs and $\tau = 0.1$.

5.5.2 Baseline Model

For comparison three supervised methods: VecMap[42], $VecMap + CL_{wt}$[1], and li et al. 2022 ($VecMap + CL_{wt} + LM/LLM + CL_{wt}$)[1] chosen as baseline models. A few-short prompting method on the llama model proposed in [18] is also chosen. The Llama 3.2 (1 Billion parameters)[96] is trained in a continued training process to incorporate all the language pairs in our experiment. We also evaluate BDI on pre-trained language model embeddings (mBERT, IndicBert for En-Mn, and mT5). The baseline and proposed methods are evaluated on Bilingual Dictionary Induction (BDI) at P@5 (Precision at 5) in CSLS (Cross-Domain Similarity Local Scaling).

5.5.3 Machine Translation

The proposed and baseline models are evaluated in a sequence-to-sequence extrinsic task: machine translation. We used fairseq sequence modeling tool[97]. We initialize the baseline[1] and propose cross-lingual embeddings to the fairseq model instead of random embedding with dimension 768. We consider 100K parallel training sentences, 1000 parallel validation sentences, and 1000 parallel testing sentences for all four language pairs. For En-Mn, the parallel sentences are extracted from the PMIndia corpus. For En-Fi, En-Tr, and En-Ta, the parallel sentences are extracted from the same dataset described in section 5.4. For machine translation, the evaluation metric is the chRF score.

5.5.4 Cross-lingual Sentence Retrival Task (CSRT)

The proposed and baseline[1] model are also evaluated in sentence level intrinsic task: Cross-Lingual Sentence Retrieval Task (CSRT). We take 10000 parallel sentences for all language pairs. Out of these 10000 sentences, the first 1000 sentences are used for testing. Cross-lingual sentence embedding is generated by taking the average of the cross-lingual word embeddings in the sentence. For En-Mn, the parallel sentences are extracted from the PMI corpus. The parallel sentences are extracted from the same dataset described in section 5.4 for En-Fi, En-Tr, and En-Ta. The baseline and proposed methods are evaluated at P@5 (Precision at 5).

5.6 Results and Discussion

5.6.1 BDI results

En-Mn. The highest increase is observed in $MACE_{MP}$ (IndicBert, edit=1) with an increase of 1.57 when compared with li et al. 2022[1]. **En-Fi**. The highest increase is observed in $MACE_{MP}(mBERT$, edit=2) with an increase of 0.86. **En-Tr**. The highest increase is observed in $MACE_{SP}(mBERT$, edit=1) with an increase of 1.31. **En-Ta**. The highest increase is observed in $MACE_{MP}(mBERT$, edit=1 and 2), with an increase of 2.49. Details of BDI results are shown in Table 5.2.

Overall, the proposed method increases the BDI performance of all four language pairs. In Source-to-Target translation, the increase in performance decreases when edit distance=3. The decrease in BDI performance worsens in En-Fi as the performance given by $MACE_{SP}$ is less than the baseline li et al. 2022[1]. This decrease in performance in edit=3 is due to the decline in the semantic similarity quality of the same root positive sample as edit distance increases. One significant finding from the experimental result is that the proposed model can increase the performance even in the case of target-to-source transformation when the same-root contrastive learning objective is focused on only target language words. This positive increase is due to symmetric contrastive $loss_{wt}$ as defined in equation 5.1. The semantic alignment enforced by $loss_{sr}$ in the target language complements the cross-lingual alignment enforced by $loss_{wt}$. From Figure 5.4, it is observed that best λ for En-Mn, En-Fi and En-Ta is 0.3. For En-Tr, the best λ is 0.4. mBERT and IndicBert outperformed mT5 in all the language pairs. The underperformance of mT5 is due to the poor contribution of $mT5 + CL_{wt}$ compared to $mBERT/IndicBert + CL_{wt}$, which indirectly suggests an over-fitting of mT5 multilingual word translation properties by $loss_{wt}$.

Table 5.2: The results of evaluation (P@5, CSLS) on BDI task with $MACE_{SP}$ (Single +ve) and $MACE_{MP}$ (multi +ve). CL_{wt}: Contrastive learning (word translation). CL_{sr}: Contrastive learning (same root-word)

Method	**En → Mn**	**Mn → En**	**En → Fi**	**Fi → En**	**En → Tr**	**Tr → En**	**En → Ta**	**Ta → En**
VecMap	51.48 ± 0.01	53.03 ± 0.06	52.31 ± 0.07	71.17 ± 0.06	60.46 ± 0.06	70.89 ± 0.06	57.40 ± 0.07	61.25 ± 0.06
mT5	00.89 ± 0.06	00.46 ± 0.06	03.97 ± 0.08	05.68 ± 0.08	14.40 ± 0.07	14.97 ± 0.06	06.97 ± 0.06	08.40 ± 0.07
IndicBERT/mBERT	00.44 ± 0.06	00.44 ± 0.06	01.31 ± 0.07	01.57 ± 0.09	13.74 ± 0.07	13.74 ± 0.07	06.54 ± 0.06	06.25 ± 0.06
li etal 2022[1] (mT5)	55.11 ± 0.06	57.68 ± 0.06	55.83 ± 0.07	72.83 ± 0.07	65.83 ± 0.06	74.11 ± 0.06	59.83 ± 0.07	63.97 ± 0.06
li etal 2022[1] (IndicBert)	60.60 ± 0.06	62.54 ± 0.06	62.31 ± 0.06	78.43 ± 0.10	67.00 ± 0.09	77.43 ± 0.10	63.25 ± 0.06	67.97 ± 0.06
li et.al 2023[18] (5-shot)	09.74 ± 0.06	58.74 ± 0.06	33.83 ± 0.06	58.17 ± 0.06	53.46 ± 0.06	70.68 ± 0.06	03.74 ± 0.06	50.68 ± 0.06
				edit=1				
$MACE_{SP}(mT5)$	55.74 ± 0.06	55.89 ± 0.06	56.89 ± 0.06	72.46 ± 0.06	66.17 ± 0.06	74.60 ± 0.10	60.60 ± 0.08	63.89 ± 0.08
$MACE_{MP}(mT5)$	56.46 ± 0.06	56.74 ± 0.06	56.31 ± 0.06	72.89 ± 0.06	65.74 ± 0.07	74.31 ± 0.07	60.76 ± 0.07	63.74 ± 0.06
$MACE_{SP}(IndicBert)$	61.76 ± 0.06	64.03 ± 0.06	63.03 ± 0.06	78.46 ± 0.06	**68.31 ± 0.09**	77.74 ± 0.09	65.31 ± 0.07	68.60 ± 0.06
$MACE_{MP}(IndicBert)$	**62.17 ± 0.07**	64.60 ± 0.03	62.17 ± 0.06	**78.74 ± 0.06**	67.46 ± 0.06	77.89 ± 0.06	**65.74 ± 0.07**	**69.74 ± 0.09**
				edit=2				
$MACE_{SP}(mT5)$	56.03 ± 0.06	55.74 ± 0.06	56.60 ± 0.06	71.46 ± 0.06	65.03 ± 0.06	74.89 ± 0.06	59.31 ± 0.07	63.17 ± 0.08
$MACE_{MP}(mT5)$	55.46 ± 0.06	55.60 ± 0.06	56.97 ± 0.06	72.22 ± 0.12	64.17 ± 0.06	73.60 ± 0.06	60.76 ± 0.07	63.17 ± 0.08
$MACE_{SP}(IndicBert)$	61.14 ± 0.01	**65.03 ± 0.07**	62.14 ± 0.09	78.03 ± 0.06	68.03 ± 0.06	77.46 ± 0.06	63.89 ± 0.06	69.03 ± 0.06
$MACE_{MP}(IndicBert)$	61.14 ± 0.01	64.00 ± 0.01	**63.17 ± 0.06**	**78.74 ± 0.06**	67.74 ± 0.09	77.31 ± 0.09	**65.74 ± 0.07**	**69.74 ± 0.09**
				edit=3				
$MACE_{SP}(mT5)$	55.31 ± 0.06	55.17 ± 0.06	56.54 ± 0.06	72.46 ± 0.06	64.31 ± 0.07	72.31 ± 0.07	59.46 ± 0.06	62.46 ± 0.06
$MACE_{MP}(mT5)$	55.74 ± 0.06	55.17 ± 0.06	55.40 ± 0.06	72.22 ± 0.12	64.31 ± 0.07	73.74 ± 0.07	59.03 ± 0.08	63.17 ± 0.08
$MACE_{SP}(IndicBert)$	62.03 ± 0.03	**65.03 ± 0.07**	62.14 ± 0.09	78.31 ± 0.07	67.89 ± 0.06	**78.03 ± 0.06**	64.17 ± 0.06	67.89 ± 0.06
$MACE_{MP}(IndicBert)$	60.89 ± 0.07	63.03 ± 0.07	61.74 ± 0.07	78.03 ± 0.06	66.74 ± 0.09	77.17 ± 0.06	62.88 ± 0.06	67.60 ± 0.06

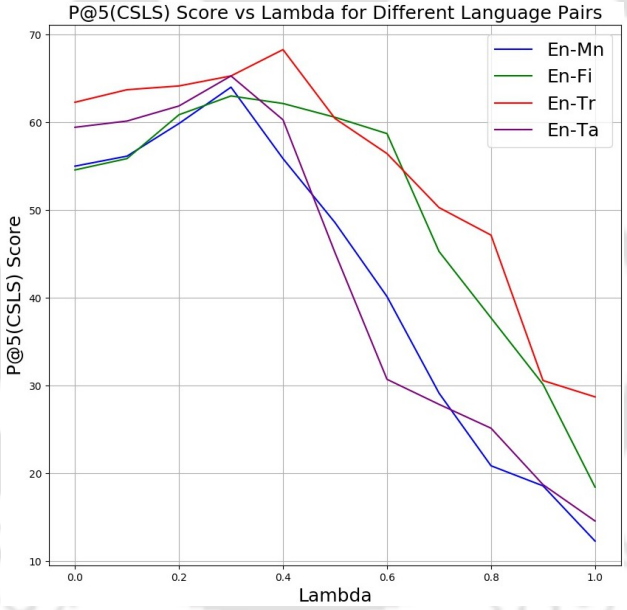


Figure 5.4: P@5 (CSLS) score on varying λ

5.6.2 mBERT/IndicBert vs mT5

The empirical analysis of Table 5.3 shows that mT5 alone outperformed mBERT/ IndicBERT (stand-alone) in all language pairs. An interesting result is observed when mBERT/IndicBert and mT5 are fine-tuned with the word translation contrastive function given in equation 5.1. mBERT/IndicBert+CL_{wt} is significantly outperforming mT5+CL_{wt} in all the language pairs. The worst case is observed in En-Fi where mT5+CL_{wt} performs less than mT5

Table 5.3: mBERT/IndicBert vs mT5 (P@5, CSLS)

Method	**En → Mn**	**En → Fi**	**En → Tr**	**En → Ta**
Indic/mBERT	00.40 ± 0.06	01.31 ± 0.06	13.17 ± 0.06	06.60 ± 0.06
mT5	00.89 ± 0.06	04.03 ± 0.06	14.46 ± 0.06	07.03 ± 0.06
Indic/mBERT+CL_{wt}	**09.03 ± 0.06**	**18.46 ± 0.05**	**28.74 ± 0.05**	**14.60 ± 0.05**
mT5+CL_{wt}	01.17 ± 0.06	03.17 ± 0.06	14.89 ± 0.06	08.17 ± 0.06
	Mn → En	**Fi → En**	**Tr → En**	**Ta → En**
Indic/mBERT	00.40 ± 0.06	01.60 ± 0.05	13.74 ± 0.06	06.31 ± 0.06
mT5	00.89 ± 0.06	05.74 ± 0.06	15.03 ± 0.06	06.31 ± 0.06
Indic/mBERT+CL_{wt}	**09.60 ± 0.06**	**23.17 ± 0.05**	**32.60 ± 0.04**	**11.74 ± 0.05**
mT5+CL_{wt}	00.89 ± 0.06	04.46 ± 0.06	15.46 ± 0.06	09.03 ± 0.06

alone. This decrease in performance observed in mT5 when fine-tuned with word translation contrastive loss ($loss_{wt}$) is most likely due to an overfitting issue. Word translation contrastive loss ($loss_{wt}$), a word-to-word association, opposes the objective function of mT5 that maximizes the probability of the correct target sequence given the input sequence. On the other hand, the objective function of mBERT/IndicBert is MLM (Mask Language Modelling) that minimizes cross-entropy loss between the predicted token (word) probabilities and the masked tokens complementing the positive word pairs and negative word pairs used in word translation contrastive loss $loss_{wt}$.

5.6.3 Machine Translation Results

Table 5.4 shows that the proposed method outperforms the baseline method in Machine Translation. The highest increase is observed in Mn-En (edit=2, $MACE_{MP}$), with an increase of 1.80. Similar to the results in BDI, the performance decreases when the edit distance becomes 3. Machine translation requires language-specific properties for sequence-to-sequence learning and is context-dependent, which is not in the scope of the proposed method. Nevertheless, the proposed method slightly outperforms the baseline model in Machine Translation.

5.6.4 Cross-lingual Sentence Retrieval Task Results

Table 5.5 shows that the proposed method outperformed the baseline model. The highest increase for the source-to-target case is observed with **MACE$_{MP}$** (edit=1) with an increase of 2.44 in **En-Mn**. The highest increase for the target-to-source case is observed with **MACE$_{MP}$** (edit=1) with an increase of 2.8 in **Ta-En**.

Table 5.4: The results of evaluation on Machine Translation tasks. (chRF score)

Method	**Mn → En**	**Fi → En**	**Tr → En**	**Ta → En**
li et al. 2022[1]	44.20 ± 0.05	44.30 ± 0.02	44.20 ± 0.05	44.20 ± 0.05
	edit=1			
$MACE_{SP}$	45.40 ± 0.02	$\mathbf{44.70 \pm 0.05}$	$\mathbf{44.80 \pm 0.02}$	$\mathbf{45.70 \pm 0.02}$
$MACE_{MP}$	45.90 ± 0.02	44.50 ± 0.05	44.50 ± 0.02	45.30 ± 0.05
	edit=2			
$MACE_{SP}$	45.20 ± 0.02	$\mathbf{44.70 \pm 0.05}$	$\mathbf{44.80 \pm 0.05}$	$\mathbf{45.70 \pm 0.02}$
$MACE_{MP}$	$\mathbf{46.00 \pm 0.02}$	44.50 ± 0.02	44.50 ± 0.05	45.30 ± 0.05
	edit=3			
$MACE_{SP}$	45.20 ± 0.05	44.30 ± 0.05	44.50 ± 0.05	45.30 ± 0.02
$MACE_{MP}$	44.50 ± 0.05	44.30 ± 0.02	44.20 ± 0.02	45.00 ± 0.02

Table 5.5: The results of evaluation (P@5) on CSRT task with $MACE_{SP}$ (Single +ve) and $MACE_MP$ (multi +ve).

Method	**En → Mn**	**Mn → En**	**En → Fi**	**Fi → En**	**En → Tr**	**Tr → En**	**En → Ta**	**Ta → En**
li et al. 2022[1]	26.78 ± 0.04	30.78 ± 0.04	29.82 ± 0.04	40.82 ± 0.04	58.98 ± 0.04	63.68 ± 0.04	53.52 ± 0.04	56.22 ± 0.04
	edit=1							
$MACE_{SP}$	29.02 ± 0.04	32.12 ± 0.04	$\mathbf{30.82 \pm 0.04}$	$\mathbf{42.52 \pm 0.04}$	$\mathbf{60.22 \pm 0.04}$	64.12 ± 0.04	53.42 ± 0.04	58.22 ± 0.04
$MACE_{MP}$	$\mathbf{29.22 \pm 0.04}$	32.52 ± 0.04	29.22 ± 0.04	40.12 ± 0.04	59.52 ± 0.04	$\mathbf{64.72 \pm 0.04}$	53.82 ± 0.04	$\mathbf{59.02 \pm 0.04}$
	edit=2							
$MACE_{SP}$	28.82 ± 0.04	30.22 ± 0.04	30.52 ± 0.04	41.82 ± 0.04	58.42 ± 0.04	63.43 ± 0.04	$\mathbf{54.02 \pm 0.04}$	57.82 ± 0.04
$MACE_{MP}$	28.52 ± 0.04	$\mathbf{32.72 \pm 0.04}$	30.02 ± 0.04	41.52 ± 0.04	59.82 ± 0.04	$\mathbf{64.72 \pm 0.04}$	53.42 ± 0.04	58.12 ± 0.04
	edit=3							
$MACE_{SP}$	28.82 ± 0.04	32.02 ± 0.04	$\mathbf{30.82 \pm 0.04}$	41.82 ± 0.04	59.82 ± 0.04	$\mathbf{64.72 \pm 0.04}$	52.71 ± 0.04	58.22 ± 0.04
$MACE_{MP}$	28.52 ± 0.04	32.12 ± 0.04	30.42 ± 0.04	42.02 ± 0.04	58.72 ± 0.04	64.22 ± 0.04	53.22 ± 0.04	58.12 ± 0.04

5.7 Choosing Best β_1, β_2 and λ

We evaluate **VecMap + CL$_{wt}$ + CL$_{sr}$** on varying β_1 (0.1 to 1) as shown in Figure 5.5. β_1 that give the best P@5 score in CSLS is chosen as the best β_1 value. Taking the best β_1, we evaluate **MACE$_{SP}$** on varying β_2 (0.1 to 1) as shown in Figure 5.5. The best β_2 is chosen similarly. The complete β_1 and β_2 values are shown in Table 5.6. Similarly as described in section 5.7, **VecMap + CL$_{wt}$ + CL$_{sr}$** is evaluated on varying β_1 (0.1 to 1) as shown in Figure 5.5. β_1 that give the best P@5 (CSLS) score is chosen as the best β_1 value. Taking the best β_1, we evaluate **MACE$_{SP}$** and **MACE$_{MP}$** on varying β_2 (0.1 to 1) as shown in Figure 5.5. The best β_2 is chosen similarly. The values of β_1 and β_2 for single +ve and multi +ve are shown in Table 5.6.

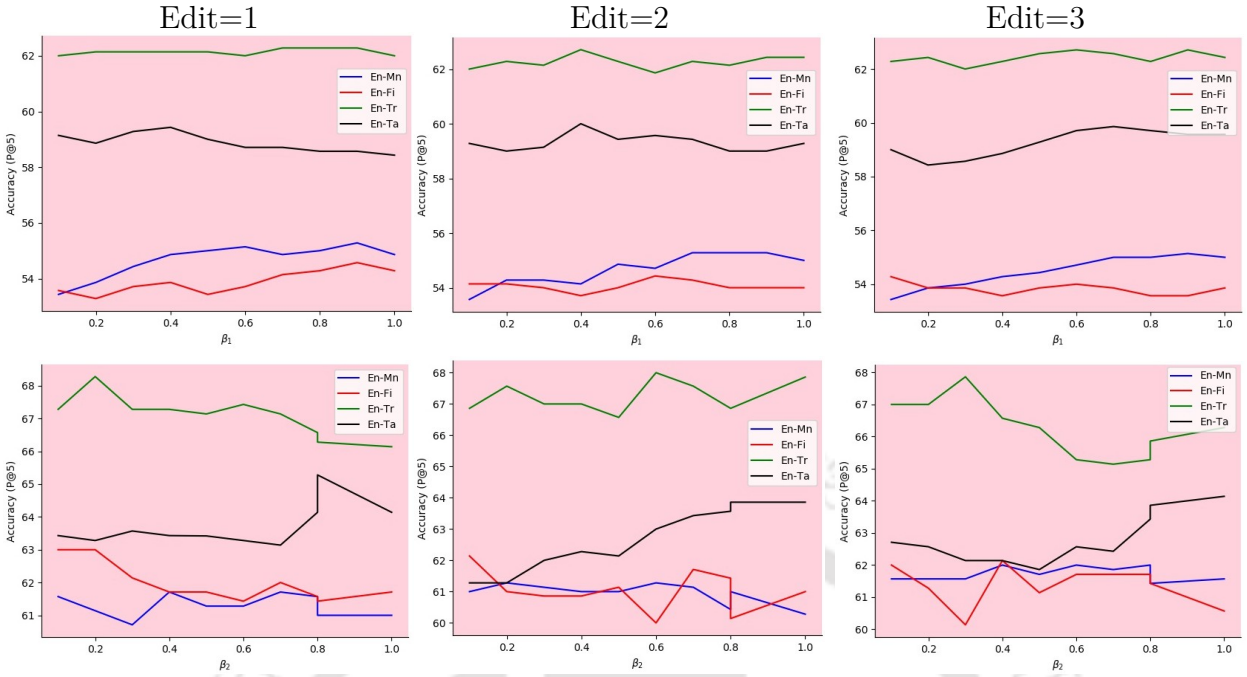


Figure 5.5: BLI scores (P@5, CSLS) on varying β_1 and varying β_2 (with best β_1)

Table 5.6: The best β_1 and β_2

scenario	Edit	parameter	En-Mn	En-Fn	En-Tr	En-Ta
Single +ve	1	β_1	0.9	0.9	0.7	0.4
		β_2	0.4	0.2	0.2	0.9
	2	β_1	0.9	0.9	0.4	0.4
		β_2	0.3	0.1	0.6	0.9
	3	β_1	0.9	0.1	0.9	0.7
		β_2	0.6	0.4	0.3	1.0
Multi +ve	1	β_1	0.8	1.0	0.4	0.4
		β_2	0.8	0.3	0.8	1.0
	2	β_1	0.8	1.0	0.4	0.4
		β_2	0.3	0.1	0.1	1.0
	3	β_1	0.9	0.1	0.4	0.7
		β_2	0.1	0.3	0.2	0.9

5.8 Ablation Study

As mBERT/IndicBert (LM) is giving better performance than mT5, we evaluate $LM + CL_{wt}$, $LM + CL_{wt} + CL_{sr}$, $VecMap + CL_{wt} + CL_{sr}$, and $VecMap + CL_{wt} + CL_{sr} + LM + CL_{wt} + CL_{sr}$ using mBERT/IndicBERT pre-trained embeddings. $LM + CL_{wt} + CL_{sr}$, $VecMap + CL_{wt} + CL_{sr}$, and $VecMap + CL_{wt} + CL_{sr} + LM + CL_{wt} + CL_{sr}$ are evaluated both in single +ve and multi +ve scenarios. Table 5.7 shows that the proposed model, $VecMap + CL_{wt} + CL_{sr} + LM + CL_{wt} + CL_{sr}$ outperformed its counterparts in 83.33% of the time (10 out of 12 cases) for Single +ve ($MACE_{SP}$) in Source-to-Target. Again, the proposed method outperformed its counterparts in 66.67% of the time (8 out of 12 cases) for Single +ve ($MACE_{SP}$) in Target-to-Source. The proposed method also outperformed its

Table 5.7: Ablation experiment results on BDI task (P@5, CSLS) with Single +ve scenario (SP) and multi +ve scenario (MP)

Method	En → Mn	Mn → En	En → Fi	Fi → En	En → Tr	Tr → En	En → Ta	Ta → En
LM+CL_{wt}	09.03 ± 0.06	09.54 ± 0.06	18.46 ± 0.06	23.11 ± 0.06	26.74 ± 0.07	32.43 ± 0.06	12.60 ± 0.06	17.03 ± 0.06
VecMap+CL_{wt}+LM+CL_{wt}	60.60 ± 0.06	62.54 ± 0.06	62.31 ± 0.06	78.43 ± 0.10	67.03 ± 0.09	77.43 ± 0.10	63.25 ± 0.06	67.97 ± 0.06
edit=1								
LM+CL_{wt}+CL_{sr} (SP)	12.31 ± 0.07	12.89 ± 0.06	19.03 ± 0.06	26.03 ± 0.06	28.03 ± 0.03	30.60 ± 0.06	14.60 ± 0.06	17.56 ± 0.06
LM+CL_{wt}+CL_{sr} (MP)	12.17 ± 0.06	12.74 ± 0.06	19.74 ± 0.06	26.46 ± 0.06	26.74 ± 0.06	32.03 ± 0.06	14.17 ± 0.06	16.60 ± 0.06
VecMap+CL_{wt}+CL_{sr} (SP)	55.31 ± 0.07	56.03 ± 0.06	54.60 ± 0.06	70.03 ± 0.06	62.31 ± 0.07	73.03 ± 0.06	59.46 ± 0.06	63.46 ± 0.06
VecMap+CL_{wt}+CL_{sr} (MP)	55.31 ± 0.07	56.17 ± 0.06	54.17 ± 0.06	71.31 ± 0.06	62.46 ± 0.06	72.74 ± 0.06	59.89 ± 0.06	63.46 ± 0.06
VecMap+CL_{wt}+CL_{sr}+LM+CL_{wt}+CL_{sr} ($MACE_{SP}$)	61.76 ± 0.06	64.03 ± 0.06	63.03 ± 0.06	**78.74 ± 0.06**	**68.31 ± 0.09**	**77.89 ± 0.06**	65.31 ± 0.07	**69.74 ± 0.09**
VecMap+CL_{wt}+CL_{sr}+LM+CL_{wt}+CL_{sr} ($MACE_{MP}$)	**62.17 ± 0.07**	64.60 ± 0.03	62.17 ± 0.06	**78.74 ± 0.06**	67.46 ± 0.06	**77.89 ± 0.06**	**65.74 ± 0.09**	**69.74 ± 0.09**
edit=2								
LM+CL_{wt}+CL_{sr} (SP)	12.03 ± 0.06	12.89 ± 0.06	19.60 ± 0.06	25.03 ± 0.06	28.31 ± 0.07	30.89 ± 0.06	14.31 ± 0.06	16.46 ± 0.06
LM+CL_{wt}+CL_{sr} (MP)	11.17 ± 0.06	12.46 ± 0.06	20.46 ± 0.06	25.74 ± 0.06	27.17 ± 0.06	31.60 ± 0.06	14.17 ± 0.06	16.60 ± 0.06
VecMap+CL_{wt}+CL_{sr} (SP)	55.31 ± 0.07	55.74 ± 0.06	54.46 ± 0.06	70.31 ± 0.06	62.74 ± 0.07	72.60 ± 0.06	60.03 ± 0.06	63.31 ± 0.06
VecMap+CL_{wt}+CL_{sr} (MP)	55.17 ± 0.06	55.74 ± 0.06	54.46 ± 0.06	70.60 ± 0.06	62.46 ± 0.06	72.74 ± 0.06	59.89 ± 0.06	63.46 ± 0.06
VecMap+CL_{wt}+CL_{sr}+LM+CL_{wt}+CL_{sr} ($MACE_{SP}$)	61.14 ± 0.01	**65.03 ± 0.07**	62.14 ± 0.09	**78.74 ± 0.06**	68.03 ± 0.06	77.31 ± 0.09	63.89 ± 0.06	**69.74 ± 0.09**
VecMap+CL_{wt}+CL_{sr}+LM+CL_{wt}+CL_{sr} ($MACE_{MP}$)	61.14 ± 0.01	64.00 ± 0.01	**63.17 ± 0.06**	**78.74 ± 0.06**	67.74 ± 0.09	77.31 ± 0.09	**65.74 ± 0.07**	**69.74 ± 0.09**
edit=3								
LM+CL_{wt}+CL_{sr}(SP)	12.31 ± 0.07	13.17 ± 0.06	19.03 ± 0.06	25.46 ± 0.06	28.31 ± 0.07	30.89 ± 0.06	14.03 ± 0.06	16.89 ± 0.06
LM+CL_{wt}+CL_{sr}(MP)	10.31 ± 0.06	11.74 ± 0.06	19.89 ± 0.06	26.74 ± 0.06	28.31 ± 0.06	32.03 ± 0.06	12.17 ± 0.06	15.89 ± 0.06
VecMap+CL_{wt}+CL_{sr} (SP)	55.03 ± 0.06	56.03 ± 0.06	54.31 ± 0.07	71.31 ± 0.06	62.74 ± 0.07	72.74 ± 0.06	59.89 ± 0.06	63.17 ± 0.06
VecMap+CL_{wt}+CL_{sr} (MP)	55.31 ± 0.06	55.31 ± 0.06	54.31 ± 0.06	71.31 ± 0.06	62.46 ± 0.06	72.60 ± 0.06	60.31 ± 0.06	63.74 ± 0.06
VecMap+CL_{wt}+CL_{sr}+LM+CL_{wt}+CL_{sr} ($MACE_{SP}$)	62.03 ± 0.03	**65.03 ± 0.07**	62.14 ± 0.09	78.03 ± 0.06	67.89 ± 0.06	77.17 ± 0.06	64.17 ± 0.06	67.60 ± 0.06
VecMap+CL_{wt}+CL_{sr}+LM+CL_{wt}+CL_{sr} ($MACE_{MP}$)	60.89 ± 0.07	63.03 ± 0.07	61.74 ± 0.07	78.03 ± 0.06	66.74 ± 0.09	77.17 ± 0.06	62.88 ± 0.06	67.60 ± 0.06

counterparts in 66.67% of the time (8 out of 12 cases) for Multi +ve ($MACE_{MP}$) in Source-to-Target. Again, the proposed method outperformed its counterparts in 66.67% of the time (8 out of 12 cases) for Multi +ve ($MACE_{MP}$) in Target-to-Source. $LM + CL_{wt} + CL_{sr}$ outperformed $LM + CL_{wt}$ in 75% of the time (12 out of 16) cases for all the language pairs when edit distance=1 (both single +ve and multi +ve). When edit=2 (both single +ve and multi +ve), $LM + CL_{wt} + CL_{sr}$ outperformed $LM + CL_{wt}$ in 68.75% of the time (11 out of 16 cases) for all the language pairs. In the case of edit=3 (both single +ve and multi +ve), $LM + CL_{wt} + CL_{sr}$ outperformed $LM + CL_{wt}$ in 62.50% of the time (10 out of 16) cases for all the language pairs. Lesser performance in edit=2 and edit=3 than edit=1 is due to poor semantic quality between the same root target words when the edit distance is larger (greater than 1). Results from Table 5.7 reconfirms the finding in [69, 85] that static cross-lingual VecMap WEs outperform LM like mBERT in BDI. The result also shows that the proposed method improves the BDI performance of mBERT/IndicBert as a stand-alone model.

5.9 t-SNE WEs visualisation plot

In monolingual fastText, cross-lingual VecMap embedding spaces, words with the same root word tend to be far apart for morphologically rich languages as visualized in Figure 5.6 and Figure 5.7. The word pair word1, word2, which have the same root word, are not close to each other, as word2s do not spread across word1s. For all the languages, the average cosine similarity between word1s and word2s in fastText, VecMap, and li et al. 2022[1] is less than the proposed method. In En-Mn, the average cosine similarity between word1s and

word2s given by the proposed method is 0.74 ± 0.43 which is greater than that of fastText (-0.36 ± 0.63), VecMap (0.06 ± 0.72) and li et al. 2022[1] (0.32 ± 0.63). The proposed method can bring words with the same root closer than fastText, VecMap, and li et al. 2022[1]. The word2s are spread across word1s significantly better as compared to fastText, VecMap, and li et al. 2022[1]. From Figure 5.6 and Figure 5.7, it is observed that, in all the languages, the proposed method is giving higher average cosine similarity between word1 and word2 as compared to fastText, VecMap, and li et al. 2022[1]. The properties of WEs gained with the proposed method help to improve the performance of BDI.

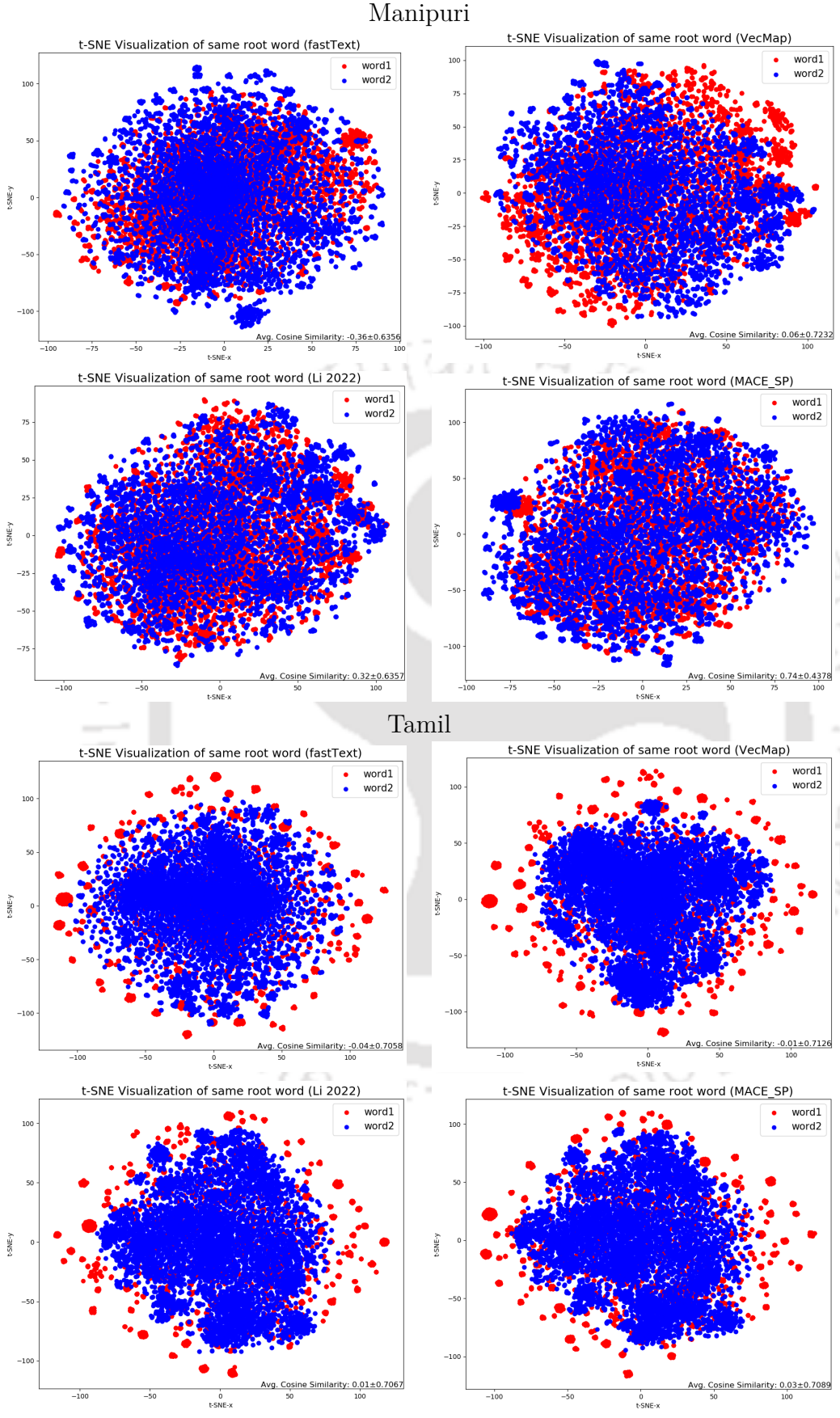


Figure 5.6: A t-SNE visualisation of words with the same root-word (word2 and word2) for Manipuri and Tamil

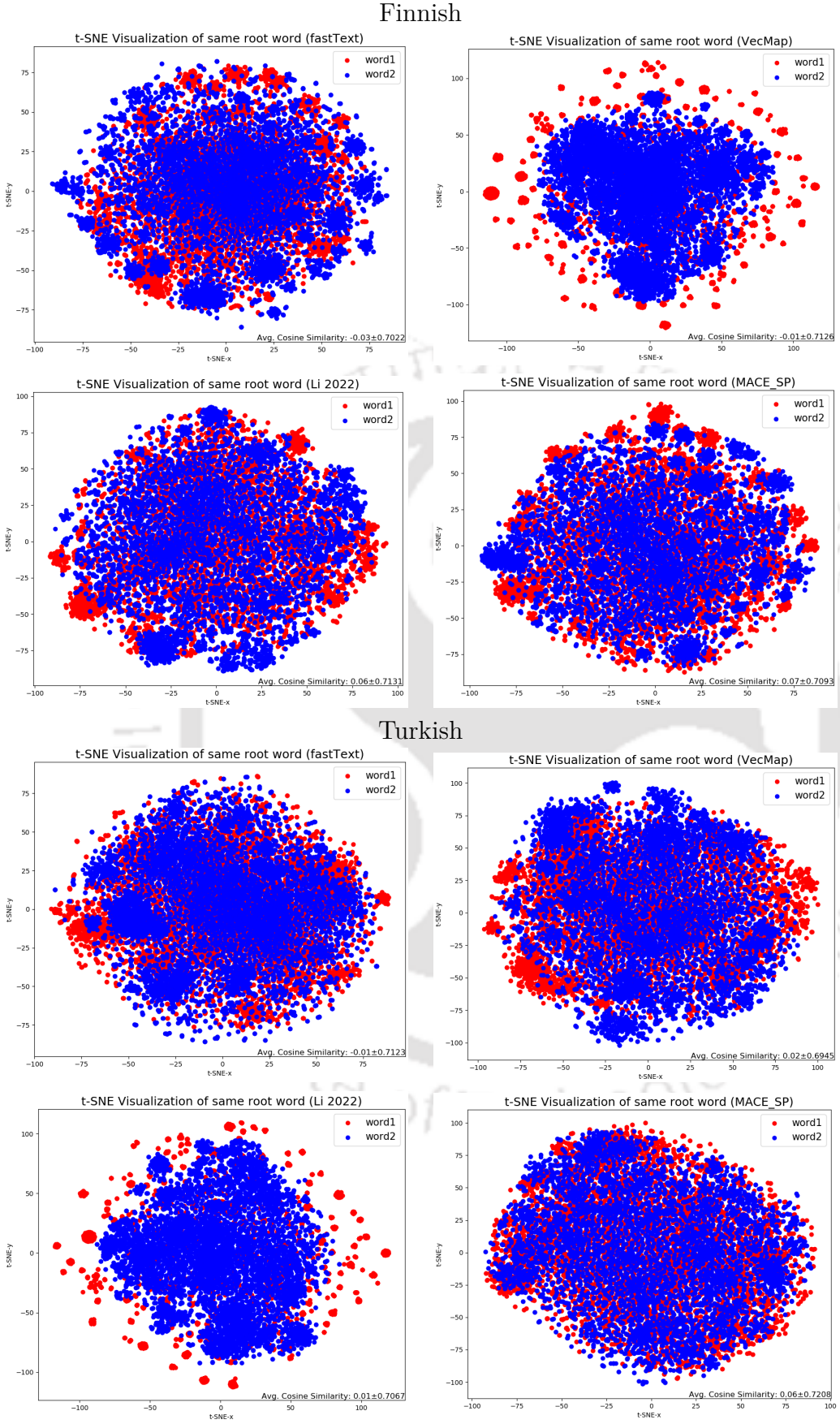


Figure 5.7: A t-SNE visualisation of words with the same root-word (word2 and word2) for Finnish and Turkish

5.10 Error Analysis

The results of li et al. 2022[1] and $MACE_{SP}$ over the BDI task are chosen for error analysis. The five nearest neighbors given by CSLS are considered for analysis. The English word, five nearest neighbor translations, the English meaning, and the ground truth Manipuri translation are given in Table 5.8. English meaning of five nearest neighbor translations are obtained using FinnWordNet, TurkishWordNet, TamilWordNet for En-Fi, En-Tr, and En-Ta respectively. For En-Mn, the author, being a native Manipuri speaker and fluent in English, translated the English word to Manipuri by himself. The baseline model fails to give the correct translation of the English words **name**, **outside**, **watching**, and **murderer**. The predicted word **মমিংদা** (in his name) and **মিংদা** (in the name) are morphological inflected with the suffix **দা** which changes the meaning slightly. In the case of the predicted word **মপান্দগী** and **মপান্দা**, these predicted words are inflected with the suffix **দগী** and **দা** respectively which changes the meaning from **"from outside"** to **"at outside"**. Similarly, the predicted words **য়েংবীয়ু** (have a look) and **মীহাৎতা** (murdered) are inflected with the suffix **বীয়ু** and **তা** respectively which changes the meaning as compared to the ground truth word **য়েংলগা** (watching) and **মীহাৎপা** (murderer) respectively. On the contrary, the proposed method can give a correct translation that the baseline method fails to do so. In the case of the English word **additional**, the proposed method fails to give the exact word translation, i.e. অমগা. Still, it gives the transliterated word এিদসেনল (additional) in the five nearest neighbors where the baseline model fails to do so. Both baseline and proposed methods fail to give the exact translations ওহৈথাকপা, মিীশং, মীচম, য়ূমশৈকশা for the English word **possibility**, **population**, **common**, and **Wednesday** respectively. Similarly, for the remaining language pairs, the strengths and weaknesses of the proposed method can be seen in Table 5.8.

Table 5.8: li et al. 2022 [1] Vs Proposed in terms of five nearest predicted translation of English word. The English meaning of the predicted words are given in the next row. *: shows strength of the proposed method and +: shows weakness of the proposed method. Bold shows correct translation

English word	li et al. 2022[1]	Proposed Predicted	Ground truth
name*	মমিংতা, মমিংদা, খনগৎলগা, থাগৎলগনি, অদি in the name, in his name, selected one, will be elected, adi	**মমিং**, মমিংদা, মিংদা, ফোতো, লাইরিকশিং name, in the name, in his name, photo, books	**মমিং**
outside*	মপান্দগী, তারগসু, লমদগী, মপান্দা, পেসেঞ্জরশিং from outside, listening, that place, at outside, passengers	মপান্দগী, লমদগী, মপান্দা, **মপান**, পেসেঞ্জরশিং from outside, that place, at outside, listening, passengers	**মপান**
watching*	য়েংবীয়ু, ইবদা, খঙঙি, নখোয়না, ঐবু have a look, in writing, I know, all of you, I	য়েংবীয়ু, ইবদা, **য়েংলগা**, ঐবু, নখোয়না have a look, in writing, watching, I, all of you	**য়েংলগা**
murderer*	ফিজম, ফরুক, ধনবির, মীহাৎতা, প্রবিস Phejam, Farukh, Dhanabir, in death, Pravish	ফিজম, **মীহাৎপা**, ধনবির, প্রবিস, মীহাৎতা Phejam, murderer, Dhanabir, Pravish, in death	**মীহাৎপা**
additional*	সেন্ট্রেল, সেন্ট্রল, বোর্ন, সেন্ট, জোতিন central, centrol, born, sent, person's name	**এদিসনেল**, এদিয়েল, এদ্দিসনেল, , এডিসনেল additional, additional, additional, additional, additional	**অমগা**
possible+	অরোৎপা, খিদোকপা, খংবদা, অকায়গী, মাঙজৌনবা secret, investigation, in knowing, for akai, beforehead	অরোৎপা, খংহন্নবা, তাংদাংবা, থেংন্নবা, খংবদা secret, disclaimer , beggar, confront, in knowing	**ওইথোকপা**
population+	পোপুলেসনগী, মীশীং, পোপুলেসন্দা, মীশীংনা, মীশীংগী for population, population, on population, population is, for population	পোপুলেসনগী, মীশীং, পোপুলেসন্দা, মীশীংনা, মীশীংগী for population, population , on population, population is, for population	**মীশিং**
common+	পাম্নগদবা, মীয়ম্না, মাম্বা, অমগা-অমগা, লৈনগদবা will be liked, people has, same, one vs one, for staying	পাম্নগদবা, ইমোসনেল, কম্মন, অমখক্তদা, কম্মন will be liked, emotional , common, only, common	**মীচম**
wednesday+	নোংমাইজং, ঙরাং, শগোলশেল, নুমিত্তা, লৈবাকপোকপা sunday, yesterday, thursday, that day, tuesday	নোংমাইজং, ঙরাং, শগোলশেল, লৈবাকপোকপা, নুমিত্তা sunday, yesterday , thursday, tuesday, that day	**য়ুমশকৈশা**
company*	yhtiö, yhtiön, yhtiössä, yrityksen, tytäryhtiön corporation, of the company, in the company, corporate, subsidiary	yhtiön, yhtiö, yhtiössä, yrityksen, **yritys** of the company, corporation, in the company, corporate, company	**yritys**
professional*	ammatillinen, ammattiin, ammattimaisuus, ammattimaisen, ammatin occupational, occupational, professionalism, corporate, profession	ammattiin, ammattimaisen, ammattimaisuus, **ammatti**, ammattimaisten occupational, corporate, professionalism, professional, profession	**ammatti**
labelling*	merkintää, merkinnän, pakkausmerkintöjen, merkintä, pakkausmerkintöjä marking, of the entry, packaging labels, entry, packaging labels	merkintää, merkinnän, merkintä, pakkausmerkintöjen, **merkintöjä** marking, of the entry, entry, packaging labels, labelling	**merkintöjä**
praise*	ylistän, kehua, kehuja, ylistää, ylistys I praise, boost, braggart, glorify, celebration	ylistän, kehuja, kehua, **kiitosta**, ylistys I praise, braggart, boost, praise, celebration	**kiitosta**
second+	kolmas, toisen, ensimmäinen, toinen, viimeinen third, another, first, other , last	kolmas, toisen, ensimmäinen, toinen, viimeinen third, another, first, other , last	**toisena**
did+	voinutkaan, olikaan, aiokaan, tekivätkin, halusi you could, it was, will, they did , wanted to	voinutkaan, olikaan, aiokaan, tekivätkin, halusi you could, it was, will, they did , wanted to	**teki**
proposed+	ehdotettu, ehdotettiin, ehdottaman, ehdottama, ehdottamaa suggested, was suggested, suggested by, suggested by , suggested	ehdotettu, ehdotettiin, ehdottaman, ehdottama, ehdottanut suggested, was suggested, suggested by, suggested by , suggested by	**ehdotetut**
itself+	vahvaksi, sellaiseksi, tällaiseksi, johtavaksi, unioninkin to be strong, to be like that, to be like this, leading , of the Union as well	sellaiseksi, tällaiseksi, johtavaksi, vahvaksi, sellaisessa to be like that, to be like this, leading, to be strong , in such	**itsessään**
included*	dahildi, dahil, dâhildir, dahildir, listelenmiştir was included, including, is included, is included, listed	dahildi, dahil, dâhildir, dahildir, **dâhil** was included, including, is included, is included, included	**dâhil**
return*	dönüşünü, dönüşünün, geri, dönülebilir, dönmek your return, of your return, back, reversible, rotate	dönüşünü, dönüşünün, **dönüş**, dönüşünüz, dönmek your return, of your return, return, your return, revolve	**dönüş**
direction*	yönün, yönüne, istikamette, yönünün, yönünde your direction, towards the direction, in the direction, direction of, towards	yönün, yönüne, istikamette, yönünün, **yön** your direction, towards the direction, in the direction, direction of, direction	**yön**
combine*	birleşmek, birleştirmek, birleştirebileceğiniz, katılmak, birleştirerek unite, to connect, you can combine, to join, by combining	birleştirin, birleştirmek, **birleştir**, birleştirebileceğiniz, birleştirir combine, to connect, combine, you can combine, combines	**birleştir**
together+	araya, arasına, gelerek, birlikte, buluşturarak between, between, by coming, with , by bringing together	araya, arasına, gelerek, birlikte, buluşturarak between, between, by coming, with , by bringing together	**beraber**
away+	uzaklıktan, mesafesinden, yürüyüş, mesafesindeler, yürüyerek from a distance, from distance, walk, they are in the distance , on foot	uzaklıktan, yürüyüş, mesafesinden, mesafesindeler, yürüyerek from a distance, walk, from distance, they are in the distance , on foot	**uzakta**
providing+	sunarak, vermektedir, sağlamaktayız, sunabilmektir, sunabilmeleri by offering, gives, we provide, is to present , be able to offer	vermektedir, sunarak, sunabilmeleri, sunabilmesi, sunabilmektir gives, by offering, be able to offer, to be able to offer , is to present	**sağlayan**
plant+	santralidir, santralın, santralinde, santrali, santralı is the power plant, of the power plant, at the switchboard, power plant , power plant	santralidir, santralın, santralinde, santrali, santralı is the power plant, of the power plant, at the switchboard, power plant , power plant	**bitki**
making*	செய்யும், செய்வதற்குத், தயார்படுத்தும், செய்வதாகும், உள்ளடக்கியுள்ளோம் will do, to do, Prepare, is to do, We have included	செய்யும், **செய்தல்**, செய்வதற்குத், செய்வதாகும், தயார்படுத்தும் will do, doing, to do, is to do, of prepare	**செய்தல்**
setup*	எக்ஸ்பெரிமெண்ட், வியூகம், உண்மையாக, செய்து, என்பதல்ல experiment, strategy, truly, done, not that	எக்ஸ்பெரிமெண்ட், செய்து, வியூகம், நிரந்தரமாக, **அமைவு** experiment, done, strategy, permanently, setup	**அமைவு**
shah*	அமித்ஷா, ரவிசங்கர், ராஜீவ், ராஜ்ய, ராஜ்நாத் amit shah, rajiv shankar, rajiv, kingdom, rajnath	அமித்ஷா, ராஜீவ், பூபேந்தர், ராஜ்நாத், **ஷா** amit shah, done, bhupender, rajnath, shah	**ஷா**
role+	வகிக்கிறது, வகித்து, வகிக்க, வகிக்கும், பங்கைப் plays, acting, to play, will play, share	வகிக்கிறது, வகிக்க, வகித்து, வகிக்கும், பங்கைப் plays, to play, acting, will play, share	**பாத்திரம்**
becomes+	ஆகிறது, மாறும், மாறுகிறது, மாறும்போது, ஆக்குகிறது is, will change, changes, when changing, makes	ஆகிறது, மாறும், மாறும்போது, மாறுகிறது, ஆக்குகிறது is, will change, when changing, changes, makes	**ஆகிறார்**
within+	இடைவெளியிலும், எல்லைக்குள், இருக்குமானால், நீடிக்கும், வரம்புக்குள் In the interval, within the limits, If there is, Lasts, within limits	இடைவெளியிலும், நீடிக்கும், இருக்குமானால், எல்லைக்குள், வரம்பும் In the interval, lasts, if there is, within the limits, limitation	**க்குள்**

5.11 Summary

The proposed method can bring words with the same root closer than fastText, VecMap, and li et al. 2022[1]. The experimental observation shows that bringing target words with the same root closer and target words with different roots apart using contrastive learning improves the BDI task's performance. Experiments on Machine Translation and Cross-lingual Sentence Retrieval Tasks also show that our proposed method outperforms the baseline method. Word embedding extracted from multilingual LM, such as mBERT/ IndicBert, doesn't have lexical properties on par with those of static word embedding, such as VecMap. Experiment results also show that the proposed method performs better in mBERT/IndicBert than mT5. The $loss_{wt}$ gives little performance increase in mT5 compared to mBERT/IndicBert due to overfitting. The ablation study reveals that the proposed contrastive learning further improves the lexical properties of mBERT/IndicBERT WEs. Different values of β_1 and β_2 for different language pairs suggest that the contribution of the proposed contrastive learning depends on the quality of the positive and negative sample pairs, which is again dependent on the language itself.

Chapter 6

Evaluation SOTA LLMs for Linguistically distant Language pairs

Bilingual Dictionary Induction (BDI) poses significant challenges in distant language pairs, especially when considering resource disparities and the complexity of linguistic structures. This chapter systematically evaluates the performance of unsupervised, supervised fine-tuning, and few-shot prompting approaches on BDI using Large Language Models (LLMs) on a diverse set of distant language pairs. The unsupervised approach explores the inherent multilingual capabilities of LLMs without fine-tuning, while the supervised fine-tuning method utilizes extensive labeled datasets to train models explicitly for BDI tasks. On the other hand, few-shot prompting leverages minimal examples to elicit accurate responses from the LLMs in a zero-shot or few-shot learning paradigm. Our experimental results reveal that the 5-shot prompting approach outperforms unsupervised and zero-shot settings in all cases and surpasses supervised settings in 82.86% of the cases. Few-shot prompting demonstrates robustness against over-fitting, leveraging LLMs' In-context learning multilingual capabilities, making it particularly effective in target-to-source translation even for morphologically complex language pairs. At the same time, few-shot prompting in LLM models like Llama is still ineffective for morphologically rich language pairs like En-Mn and En-Ta in source-to-target BDI tasks. These findings suggest few-shot prompting as a cost-effective and powerful alternative for BDI tasks, with future work focusing on prompt optimization.

6.1 Introduction

Bilingual Dictionary Induction (BDI) serves as a foundational task in natural language processing (NLP), enabling the automatic creation of bilingual dictionaries by aligning word

representations across languages[7]. This capability is critical for cross-linguistic resource development, particularly for applications such as machine translation[61] and cross-lingual information retrieval[60]. A widely adopted approach to BDI is cross-lingual mapping[8], where word embeddings from different languages are transformed into a shared vector space, facilitating the alignment of lexical resources. Among these methods, matrix factorization techniques, such as VecMap[3], have demonstrated strong performance for resource-rich language pairs, such as English-Italian (En-It). These methods rely on iterative refinement mapping[23] using Singular Value Decomposition (SVD) to improve the alignment quality. However, their performance deteriorates for resource-poor and linguistically distant[13] language pairs, such as English-Manipuri (En-Mn)[33, 34], due to challenges like data scarcity, orthographic differences, and significant linguistic divergence. To further boost the BDI performance, contrastive learning cross-lingual embedding approach[1] has emerged as a powerful approach that combines static word embeddings (VecMap) with contextual embeddings generated by language models (LMs) and large language models (LLMs). Using a contrastive framework and leveraging dictionary pairs for alignment, this method enhances performance by effectively integrating the strengths of both static and contextual embeddings. The contrastive learning approach offers a promising pathway for improving BDI across a broader range of language pairs. However, the contrastive learning approach gives lesser performance for linguistic distant and morphologically complex language pairs like English-Manipuri (En-Mn)[33, 34], English-Finnish (En-Fi)[1], and English-Turkish (En-Tr)[1] than similar language pairs like English-Italian (En-It).

Recent research has explored the potential of large language models (LLMs) to induce BDI using a few-shot prompting approach[18]. This technique involves leveraging the inherent contextual understanding of LLMs to align embeddings with minimal supervision[44], often achieving better results than traditional VecMap and contrastive learning methods. However, this study has primarily focused on a limited set of resource-rich and structurally similar (closer) language pairs, often neglecting the evaluation of linguistically challenging and distant language pairs such as English-Manipuri (En-Mn), English-Finnish (En-Fi), English-Turkish (En-Tr), English-Hindi (En-Hi), English-Japanese (En-Ja), and English-Tamil (En-Ta). Moreover, evaluation in state-of-the-art LLM models like LLaMA remains unexplored mainly in the above-mentioned challenging settings[18]. Most of the LLM models have inherent multilingual properties[98] that can also be exploited without in-context learning in an unsupervised way for BDI. Apart from the unsupervised approach, two predominant paradigms have emerged in leveraging LLMs for bilingual dictionary induction: few-shot prompting[18] and supervised fine-tuning using contrastive learning[1]. Few-shot prompting exploits the in-context learning capabilities of LLMs[19], requiring minimal task-specific examples to guide the model. While few-shot prompting is less resource-intensive[99][100], its

viability for morphologically complex and distant language pairs still needs to be explored. On the other hand, the supervised fine-tuning approach[1] involves extensive bilingual dictionary pairs and a much higher computational cost to fine-tune the model for the BDI task. While both methods hold promise, their comparative effectiveness and practical implications remain under-explored, particularly in distant linguistic pairs and morphologically challenging environments. Motivated by the concerns mentioned above, this chapter investigates the feasibility of few-shot prompting in distant and morphologically challenging language pairs, compared to supervised fine-tuning methods that demand large dictionary pairs and high computational overhead. This study provides a comprehensive evaluation of LLMs, focusing on key questions such as: *Is few-shot prompting a feasible alternative to supervised fine-tuning methods for distant and morphologically challenging language pairs, given the limited availability of dictionary pairs and the computational cost associated with supervised learning*?

6.1.1 Contribution

The major contributions of this chapter are:

(i) The chapter presents a first-ever comparative analysis of BDI performance using large language models across unsupervised, supervised fine-tuning, and few-shot settings, examining their applicability and limitations in linguistically challenging and distant language pairs.

(ii) This chapter shows that the 5-shot prompting approach outperforms unsupervised and zero-shot settings and surpasses supervised settings in 82.86% of the evaluation cases. Experimental results also show that few-shot prompting in LLM models like Llama is still ineffective for morphologically rich language pairs like En-Mn, En-Fi, En-Tr, and En-Ta in source-to-target BDI tasks.

6.2 RELATED WORK

Previous works on bilingual dictionary induction (BDI) encompass various approaches and methods. The initial mapping-based cross-lingual word embedding (CLWE) model [72] introduced a regression-based framework to learn a linear mapping function using a bilingual seed dictionary. Following this, a matrix factorization approach[3] and its variant [42] proposed a closed-form solution commonly referred to as VECMAP.[1] However, both VecMap

[1] https://github.com/artetxem/vecmap

and a centrality-aligned ridge regression-based orthogonal mapping [89] struggled to handle morphologically complex words in BDI tasks. An empirical investigation in [76] demonstrated that both supervised and unsupervised methods underperform for morphologically rich languages, such as Finnish and Turkish, in BDI tasks. Morpheme-based approaches that segment words into their root and suffix components were proposed to address this, leading to slight improvements in BDI performance [27, 28]. Further, the work in [47] introduced 40 morphologically complete dictionaries and highlighted the severe degradation in BDI performance for less frequent inflected words. Later, the method proposed in [90] used a morphologically aware probabilistic model that jointly models lexeme translation and inflectional morphology. With the rise of contrastive learning techniques, methods for contrastive learning-based cross-lingual word representation [1] was developed. However, the contrastive learning approach [1] showed limitations in handling morphologically rich languages like Finnish and Turkish. More recently, a method that leverages LLM proposed a few-shot prompting approach[18], achieving state-of-the-art BDI scores across many language pairs. However, the few-shot prompting method evaluates only resource-rich and linguistically closer language pairs, neglecting morphologically complex and distant language pairs.

6.3 Methodology

We consider two pre-trained Language Models (LM): mBERT[55] and IndicBERTv2-MLM-only[57]. We further take seven multilingual Large Language Model (LLM): $mT5_{small}$[87], $mT5_{base}$[87], $ByT5_{base}$[101], $XGLM_{564M}$[102], $mBART_{large}$[103], $mGPT_{1.3B}$[104], and $Llama-3.2_{1B}$[96]. For contrastive fine-tuning of LM/LLM, we used pre-trained $mBERT_{base}$ (En-It, En-Fi, En-Hi, En-Tr, En-Ja, and En-Ta)[55] and IndicBERTv2-MLM-only (En-Mn)[57]. $mT5_{small}$, $mT5_{base}$, $ByT5_{base}$, $XGLM_{564M}$, $nBART_{large}$, $mGPT_{1.3B}$, and $Llama-3.2_{1B}$, are trained in continue training approach[88] to incorporate Manipuri data. $Llama-3.2_{1B}$ is also trained in the continued training process[88] to incorporate Finnish, Turkish, Japanese, and Tamil languages. The details of the LLM models are shown in Table 6.1.

Three main approaches have emerged for leveraging LLMs in BDI: Unsupervised, Supervised fine-tuning[1], and few-shot prompting[18]. Supervised fine-tuning relies on large dictionary pairs and incurs significantly higher computational costs to tailor the model to specific tasks. Unsupervised approach doesn't require dictionary pairs for leveraging LLMs in BDI. While few-shot prompting utilizes the in-context learning abilities of LLMs [19], requiring only minimal task-specific examples. While all the three methods show potential, their comparative comparatively analysis, advantages, and trade-offs, particularly across diverse

Table 6.1: Model Details

LM/LLM	No of Parameters
mBERT	110 Millions
IndicBERTv2-MLM-only	278 Milliions
mT5-small	300 Millions
mT5-base	580 Millions
byT5-base	580 Millions
XGLM	564 MILLIONS
mBART-large	610 Millions
mGPT	1.3 Billions
Llama-3.2	1 Billion

language pairs and complex linguistic nature, remain under-explored. The above mentioned LM/LLM models are evaluated in supervised, unsupervised, and few-shot settings as discuss below.

6.3.1 Supervised Fine-tuning

If $z = f(x)$ defines a target language dictionary word of a source language dictionary word x, a bilingual dictionary pair is defined as $D = \{(x, z)|x \in X, z \in Z, z = f(x)\}$ where X and Z are the words in source and target languages, respectively. Given a pair $(x_i, z_i^+)\epsilon D$, x_i and z_i^+ are tokenized using mBERT tokenizer, giving the following sub-words sequences: $s_1 x_i ....s_n x_i$, $s_1 z_i^+ ....s_n z_i^+$, $n \geq 1$. The LM/LLM encoding function f_θ takes the sequence as input and gives the average representation of the token in the last transformer layer as the representation of x_i and z_i^+ respectively[69].

For supervised fine-tuning using contrastive learning, positive samples present in D and negative samples not present in D are required. Like in [1], for a given positive pair $(x_i, z_i^+)\epsilon D$, a hard negative set $S_z^- = \left\{z_j \mid (x_i, z_i^+)\epsilon D, z_j \neq z_i^+, z_j \notin NN(x_i)\right\}$ is generated where $NN(x_i)$ is the nearest neighbors of x_i from VecMap embedding of target language excluding z_i^+. Similarly, for target-to-source translation, we generate the negative pair set $S_x^- = \left\{x_j \mid (x_i, z_i^+)\epsilon D, x_j \neq x_i, x_j \notin NN(z_i^+)\right\}$ where $NN(z_i^+)$ is the nearest neighbors of z_i^+ from VecMap embedding of source language excluding x_i.

For supervised fine-tunning, we used the state-of-the-art contrastive fine-tuning approach[1] using negative pair set S_z^- and S_x^- as described above.

$$loss_{wt} = \frac{sim(x_i, z_i^+)}{\sum\limits_{z_j \epsilon \left\{z_i^+\right\} \cup S_z^-} sim(x_i, z_j) + \sum\limits_{x_j \epsilon S_x^-} sim(x_j, z_i^+)} \tag{6.1}$$

$$sim(x_i, z_j) = \exp^{\cos(f_\theta(x_i),\, f_\theta(z_j))/\tau} \tag{6.2}$$

The final contrastive learning objective function that fine-tune LM/LLM parameters θ is given in equations 6.3

$$\min_{\theta} - \left[\mathbb{E}_{(x_i, z_i^+) \in D_{wt}} \log(\text{loss}_{wt})\right] \tag{6.3}$$

6.3.2 Unsupervised Setting

In an unsupervised setting, the off-the-self average representation of the token (words) in the last transformer layer, as mentioned above, without fine-tuning, is taken for BDI evaluation.

6.3.3 Few-shot Prompting

We used the method that leverages autoregressive LLM for few-shot prompting as proposed in [18]. We perform both zero-shot and few-shot prompting using the in-context learning capabilities of LLMs[19]. For the prompting process (zero-shot and 5-shot), we used the best 'mask-filling-style' and 'GPT-style' templates as proposed in [18]. The template details are given in Table 6.2.

6.4 DATASET

For continued training process[88], this study considers five language pairs: English-Manipuri (En-Mn), English-Finnish (En-Fi), English-Turkish (En-Tr), English-Tamil (En-Ta), and English-Japanese (En-Ja). For En-Fi, the Europarl[2] parallel corpus [54] extracted from the proceedings of the European Parliament is used. Parallel corpus provided in MaCoCu-tr-en 2.0[94] is used for En-Tr. En-Ta parallel corpus from Bharat Parallel Corpus Collection (BPCC), AI4BHARAT[3] is used. For En-Ja, a parallel sub-title corpus [81] extracted from the conversational dialogue is used for English-Japanese. Finally, for En-Mn, the comparable corpus used in [33, 89, 105] are used. This study considers the bilingual dictionary available at Directorate of Language Planning and Implementation, Government of Manipur [4] for En-Mn language pair, and the MUSE[5] library for the En-It, En-Tr, En-Fi, En-Hi, En-Ja, and En-Ta language pairs. The dataset details are given in Table 6.3.

[2] https://www.statmt.org/europarl/
[3] https://ai4bharat.iitm.ac.in/bpcc/
[4] https://www.dlpi.mn.gov.in/en/
[5] https://github.com/facebookresearch/MUSE

Table 6.2: Best Template for Few-shot prompting

LLM	**Zero-shot**
$mT5_{small}$	The word 'x' in L_z is: <mask>.
$mT5_{base}$	Translate the word 'x' into L_z: <mask>.
$XGLM_{564m}$	The L_x word x in L_Z is:
$mGPT_{1.3B}$	Translate the L_x word x into L_y:
$Llama-3.2_{1B}$	The L_x word x in L_z is:
LLM	**5-shot**
$mT5_{small}$	The word x_1 in L_z is z_1. The word x_2 in L_z is z_2. The word x_3 in L_z is z_3. The word x_4 in L_z is z_4. The word x_5 in L_z is z_5. The word x in L_z is <mask>.
$mT5_{base}$	The word x_1 in L_z is z_1. The word x_2 in L_z is z_2. The word x_3 in L_z is z_3. The word x_4 in L_z is z_4. The word x_5 in L_z is z_5. The word x in L_z is <mask>.
$XGLM_{564M}$	The word x_1 in L_z is z_i. The word x_2 in L_z is z_2. The word x_3 in L_z is z_3. The word x_4 in L_z is z_4. The word x_5 in L_z is z_5. The word x in L_z is.
$mGPT_{1.3B}$	The L_x word x_1 in L_z is z_1. The L_x word x_2 in L_z is z_2. The L_x word x_3 in L_z is z_3. The L_x word x_4 in L_z is z_4. The L_x word x_5 in L_z is z_5. The L_x word x in L_z is
$Llama-3.2_{1B}$	The L_x word 'x_1' in L_z is z_1. The L_x word 'x_2' in L_z is z_2. The L_x word 'x_3' in L_z is z_3. The L_x word 'x_4' in L_z is z_4. The L_x word 'x_5' in L_z is z_5. The L_x word x in L_z is

Table 6.3: Statistics of data, **LP**: Language Pairs

LP	Platform	sentences		words		unique words	
En-Mn	**Sangai Express**	**En**	**Mn**	**En**	**Mn**	**En**	**Mn**
	+Poknafam+PMI	129546	181553	3.5M	3.3M	15247	24449
En-It	**European Parliament**	**En**	**It**	**En**	**It**	**En**	**It**
		1.90 M	1.90M	49.6M	47.4M	151,017	219,976
En-Fi	**European Parliament**	**En**	**Fi**	**En**	**Fi**	**En**	**Fi**
		1.92 M	1.92M	47.4M	32.2M	151017	219976
En-Hi	**CILT,IIT Bombay**	**En**	**Hi**	**En**	**Hi**	**En**	**Hi**
		1.6M	1.6M	23.8M	24.6M	238,765	392,634
En-Ja	**opensubtitles.org**	**En**	**Ja**	**En**	**Ja**	**En**	**Ja**
	kitsunekko.net	2.8 M	2.8M	23.6M	21.5M	154,276	138,487
En-Ta	**AI for Bharat**	**En**	**Ta**	**En**	**Ta**	**En**	**Ta**
		442776	442776	10.3M	8.4M	79518	314452
En-Tr	**MaCoCu-tr-en 2.0**	**En**	**Tr**	**En**	**Tr**	**En**	**Tr**
		1.6 M	1.6M	55.0M	51.5M	411397	884161

6.5 Experimental Setup

6.5.1 Contrastive learning parameter

The hyper-parameter values are N_{iter}=5, N_{neg}=50, N_{iter} is the no of iterations in VecMap. N_{neg} is the no of negative samples for a positive pair. AdamW[56] optimizer with a learning rate of 2e-5 is used for fine-tuning mBERT/IndicBert, $xglm_{564M}$, $mbart_{large}$, $mGPT_{1.3B}$, and $Llama-3.2_{1B}$. For fine-tuning $mT5_{small}$, $mT5_{base}$, and $byT5_{base}$, a learning rate of 6e-5 is used . LM/LLM WEs are fine-tuned for five epochs with $\tau = 0.1$.

6.5.2 Few-shot prompting

We consider zero-shot to 10-shot prompting. Like in [18], we set the beam size to 5 for all LLMs. For encoder-decoder models, the maximum sequence length is fixed at 5. In contrast, decoder-only models are set to 5 plus the input sequence length, as they first replicate the input before generating new content. For encoder-decoder LLMs, the evaluation batch size is set to 100 for smaller and 8 for larger models.

6.5.3 BDI Evaluation

The LM/LLM models are evaluated on Bilingual Dictionary Induction (BDI) at P@5 (Precision at 5). A training dictionary of 3500 and 700 testing pairs is used for all language

	LM/LLM	En → Mn	Mn → En	En → It	It → En	En → Fi	Fi → En	En → Hi	Hi → En	En → Tr	Tr → En	En → Ja	Ja → En	En → Ta	Ta → En
Unsupervised	mBERT/IndicBert	00.44	00.44	19.71	18.43	01.31	01.57	01.71	01.32	13.74	13.74	04.00	03.22	06.54	06.25
	$mT5_{small}$	**00.89**	**00.46**	21.00	22.71	**03.97**	**05.68**	01.43	01.18	**14.40**	14.97	02.14	02.78	**06.97**	**08.40**
	$mT5_{base}$	00.14	00.28	10.28	10.28	00.57	00.57	00.86	00.88	12.43	12.57	03.00	03.37	06.14	06.28
	$byT5_{base}$	00.14	00.00	18.00	27.00	01.43	03.00	00.43	00.44	12.14	**15.28**	01.71	01.76	05.57	05.57
	$xglm_{564M}$	00.14	00.00	14.28	14.43	00.86	01.00	00.43	00.44	12.43	12.43	01.71	01.76	06.14	06.14
	$mbart_{large}$	00.57	00.43	17.71	18.00	01.57	01.71	**02.14**	**02.21**	13.14	12.86	05.43	03.08	06.43	06.86
	$mGPT_{1.3B}$	00.28	00.14	**25.71**	**28.71**	01.28	01.86	00.43	00.44	13.86	13.14	**06.28**	**07.62**	06.14	06.14
	$Llama3.2_{1B}$	00.00	00.00	15.86	17.14	01.00	01.14	00.43	00.44	13.28	13.43	01.71	03.22	06.14	06.16
Supervised	mBERT/IndicBert	09.03	09.54	61.28	65.86	**18.46**	**23.11**	12.86	17.26	26.74	**32.43**	25.28	26.39	12.60	17.03
	$mT5_{small}$	07.86	09.43	41.14	47.28	12.57	16.86	05.14	07.82	21.28	24.00	30.14	29.91	15.43	16.71
	$mT5_{base}$	**36.14***	**35.43**	49.71	56.71	01.14	00.71	00.86	00.59	12.00	12.00	44.57	42.81	21.43	23.00
	$byT5_{base}$	00.57	01.00	21.00	33.71	02.57	04.57	00.43	00.44	12.71	17.00	02.14	02.49	05.57	05.57
	$xglm_{564M}$	00.71	00.71	40.14	45.43	14.71	18.57	13.14	12.83	24.86	24.43	04.57	06.01	13.28	22.71
	$mbart_{large}$	14.57	16.00	**63.57**	**66.86**	11.43	09.43	**41.28**	**42.77**	16.86	16.14	41.57	35.78	**28.86**	**30.70**
	$mGPT_{1.3B}$	00.14	00.00	40.00	43.28	02.71	03.14	00.28	00.29	18.43	18.28	32.86	29.62	06.14	06.14
	$Llama3.2_{1B}$	23.28	28.57	44.86	48.14	10.86	11.71	26.86	30.09	**30.28**	30.86	**48.14**	**43.11**	06.28	07.14
0-shot	$mT5_{small}$	00.00	00.28	09.14	20.43	00.71	02.43	00.43	01.91	11.57	13.86	01.71	13.49	**06.14**	06.86
	$mT5_{base}$	00.00	**03.57**	25.28	28.43	03.28	07.43	00.43	14.90	12.41	**17.86**	01.57	**41.35**	05.28	14.28
	$xglm_{564M}$	00.28	00.57	22.28	24.14	07.43	09.28	00.28	09.59	12.71	14.57	04.57	11.43	03.57	07.57
	$mGPT_{1.3B}$	00.00	02.43	36.57	39.28	**08.28**	**14.43**	07.71	19.47	**20.14**	13.43	**47.43**	28.15	04.43	**20.00**
	$Llama3.2_{1B}$	**05.71**	01.57	**48.00**	**53.00**	00.14	03.71	**26.00**	**56.19**	03.86	15.43	00.28	09.38	00.14	01.57
5-shot	$mT5_{small}$	05.71	08.28	38.86	56.43	14.57	33.57	16.57	25.37	26.28	44.43	37.71	46.63	17.57	33.00
	$mT5_{base}$	**14.43**	32.43	59.00	73.28	**32.00**	55.14	36.14	56.78	42.86	61.28	52.86	65.39	**36.28**	36.17
	$xglm_{564M}$	07.43	11.57	40.14	57.71	29.57	52.28	32.71	44.10	35.71	50.86	27.28	40.47	25.57	33.86
	$mGPT_{1.3B}$	08.00	20.71	64.00	**83.71**	31.14	**60.86**	21.43	56.34	**54.57**	**76.43**	**62.71**	**77.71**	13.28	**45.57**
	$Llama3.2_{1B}$	09.71	**58.71***	**67.14***	80.28	02.43	19.28	**46.71**	**65.63**	21.43	36.28	01.43	33.87	00.57	18.14

Table 6.4: The results of evaluation (P@5) on BDI task over LM/LLM

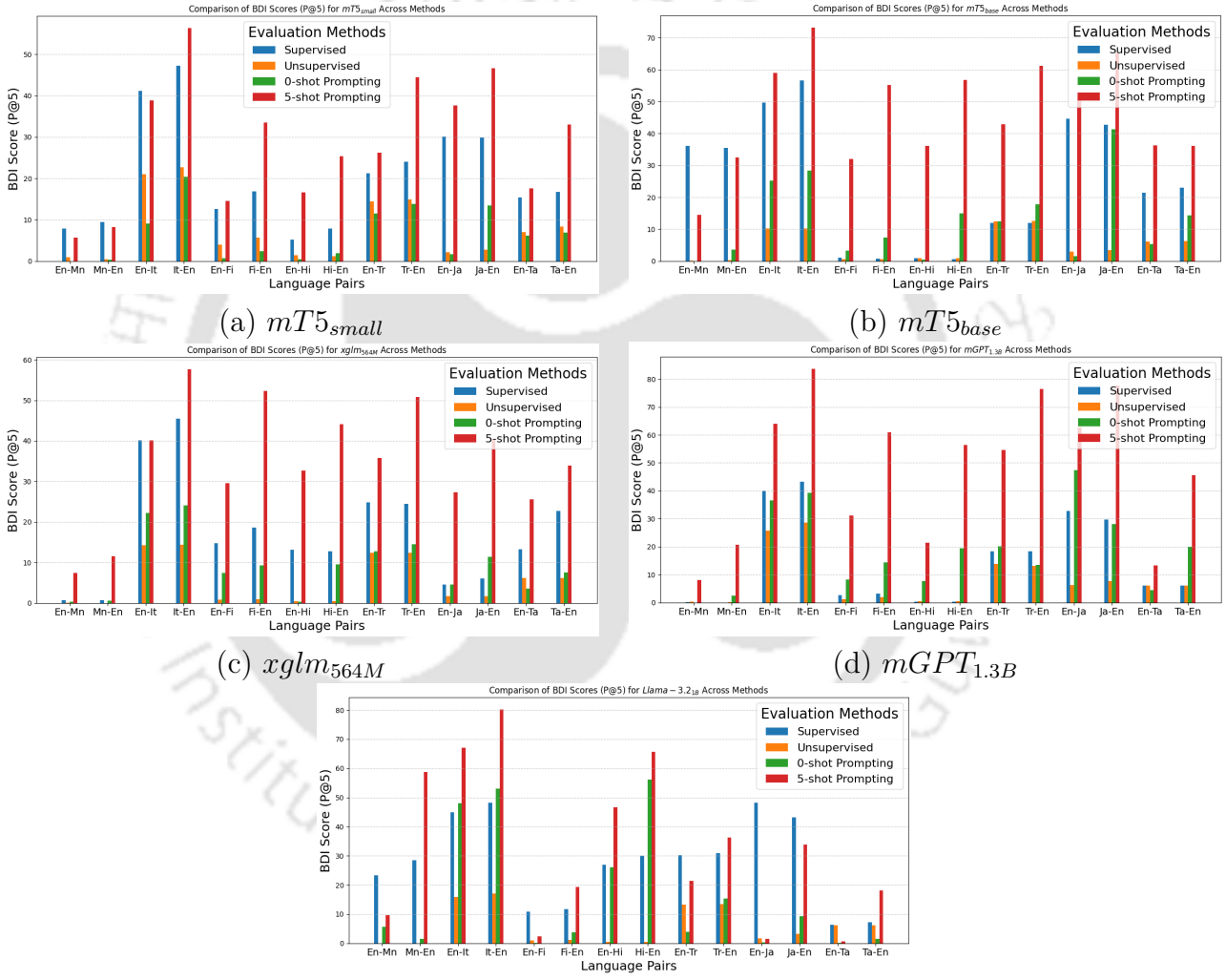
pairs.

6.6 Results and Discussion

Unsupervised: For En-Mn, Mn-En, En-Fi, Fi-En, En-Ta, and Ta-En $mT5_{small}$ gives the highest score of 00.89, 00.46, 03.97, 05.68, 06.97, and 08.40 respectively. $mGPT_{1.3B}$ gives a better score of 25.71, 28.71, 06.28, and 07.62 for En-It, It-En, En-Ja, and Ja-En, respectively. In En-Hi and Hi-En, the highest performance is shown by $mBART_{large}$ with 02.14 and 02.21, respectively. In the case of En-Tr, $mT5_{small}$ produces a higher performance score of 14.40. However, $ByT5_{base}$ performs better in Tr-En with a score of 15.28. Although no particular LLM model outperforms in all language pairs, all the LLM models perform less in linguistic distant and morphologically complex pairs than linguistically similar pairs like En-It in unsupervised settings. Details of BDI results are shown in Table 6.4.

Supervised: For En-Mn and Mn-En, $mT5_{base}$ gives higher score of 36.14 and 35.43 respectively. $mBART_{large}$ gives the highest score of 63.57, 66.86, 41.28, 42.77, 28.86, and 30.70 in En-It, It-En, En-Hi, Hi-En, En-Ta, and Ta-En, respectively. For En-Fi and Fi-En LM models, mBERT performs better than all the LLM models. In the case of En-Tr, the highest score is given by $Llama-3.2_{1B}$ with a score of 30.28, but mBERT gives the highest score of 32.43 in Tr-En. $Llama-3.2_{1B}$ gives the highest performance in En-Ja and Ja-En with a score of 48.14 and 43.11, respectively. Like in an unsupervised setting, there is no particular LLM model that outperforms in all language pairs, but all the LLM models give much lesser performance in linguistically distant and morphologically complex pairs as compared to linguistically similar pairs like En-It in an unsupervised setting. Details of BDI results are shown in Table 6.4.

Zero-shot: For En-Mn, $Llama-3.2_{1B}$ gives the highest performance score of 05.71, but $mT5_{base}$ gives best score of 03.57 in Mn-En. $Llama-3.2_{1B}$ gives the highest En-It, It-En, En-Hi, Hi-En score. For En-Fi and Fi-En, $mGPT_{1.3B}$ outperformed all the remaining LLM models. In the case of En-Tr and En-Ja, $mGPT_{1.3B}$ gives the highest score. $mT5_{base}$ outperform the remaining LLM models in Tr-En and Ja-En. For En-Ta, $mT5_{small}$ gives the highest score. On the other hand, $mGPT_{1.3B}$ gives the best score in Ta-En. Similar to unsupervised and supervised, all the LLM models perform much less in linguistic distant and morphologically complex pairs than in linguistically similar pairs like En-It in unsupervised settings. Details of BDI results are shown in Table 6.4.



(a) $mT5_{small}$ (b) $mT5_{base}$

(c) $xglm_{564M}$ (d) $mGPT_{1.3B}$

(e) $Llama-3.2_{1B}$

Figure 6.1: Supervised vs Unsupervised vs 0-shot vs 5-shot

5-shot: For En-Mn $mT5_{small}$ gives the highest performance score of 14.43, but $Llama-3.2_{1B}$ gives best score of 58.71 in Mn-En. $Llama-3.2_{1B}$ gives the highest En-It, En-Hi, and Hi-En score. $mGPT_{1.3B}$ outperformed all the remaining LLM models in It-En, Fi-En, En-Tr, Tr-En, En-Ja, Ja-En, and Ta-En. In the case of En-Fi and En-Ta, $mT5_{base}$ gives the highest

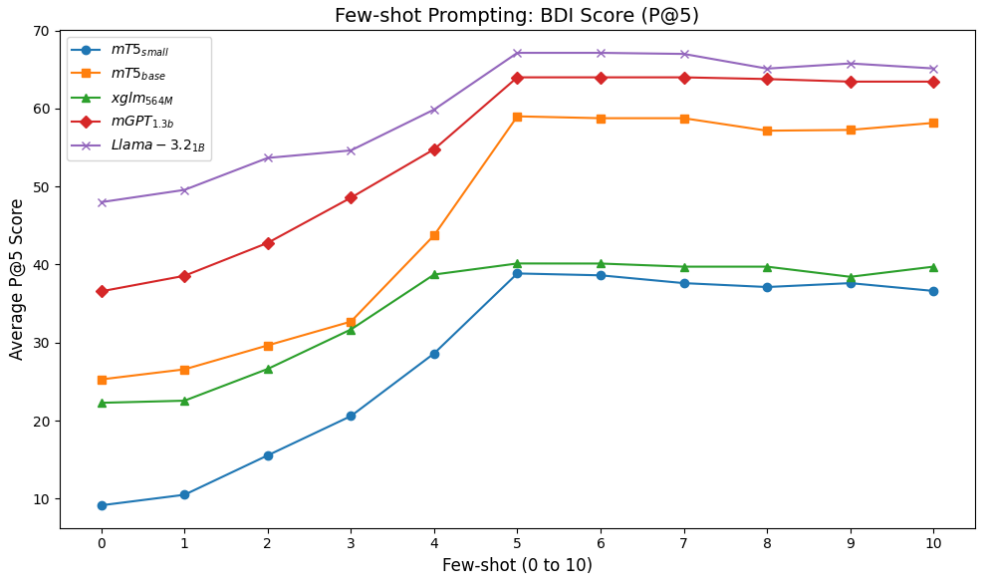


Figure 6.2: BDI scores averaged over 14 BDI directions with respect to n-shot (0 to 10)

score. $Llama-3.2_{1B}$ and $mGPT_{1.3B}$ outperform all the LLM models in 78.57% (11 cases out of 14 cases) of the time. Similarly, the 5-shot approach performs less in linguistic distant and morphologically complex language pairs like En-Mn, Mn-En, En-Ta, and Ta-En. Details of BDI results are shown in Table 6.4.

Supervise Fine-tuning vs Unsupervised vs 0-shot vs 5-shot: Fig 6.1, shows that 5-shot prompting outperformed both unsupervised and zero-shot prompting 100% (in all cases) of the time. 5-shot prompting outperforms supervised setting in 82.86% (58 cases out of 70 cases) of the time. Supervised setting outperforms 5-shot prompting in $mT5_{small}$, $mT5_{base}$, and $Llama-3.2_{1B}$ for En-Mn. For Mn-En and Ta-En, the 5-shot prompting approach outperforms supervised settings in $Llama-3.2_{1B}$. The performance differences in Mn-En/Ta-En and En-Mn/En-Ta tasks using Llama models can be attributed to the model's bias toward high-resource languages like English. While effective for En-Mn and En-Ta, supervised fine-tuning approaches may over-fit the training data, reducing their ability to generalize to unseen examples. In contrast, few-shot prompting leverages the model's inherent multilingual and contextual understanding, avoiding over-fitting and preserving the pretrained multilingual alignment. This makes the few-shot prompting approach results higher than supervised in the case of target-to-source translation: Mn-En and Ta-En.

Morphological Complex Settings: On the other hand, the supervised fine-tuning approach outperforms 5-shot prompting for En-Mn, En-Fi, En-Tr, and En-Ta in $Llama-3.2_{1B}$. The nature of the task in En-Mn, En-Fi, En-Tr, and En-Ta translation are inherently more complex as Manipuri, Finnish, Turkish, and Tamil are morphologically rich languages, making fine-tuned models more advantageous. In Mn-En, Fi-En, Tr-En, and Ta-En, the BDI task

is often simpler finding a direct English equivalent for a Manipuri/Finnish/Turkish/Tamil word), making it more amenable to few-shot prompting. This observation suggests that few-shot prompting in the Llama model is still ineffective for morphologically rich language pairs in the BDI task.

Best n-shot: For all the language pairs in our study, BDI for 0 to 10 shots are evaluated. Average BDI score over 14 BDI directions for $mT5_{small}$, $mT5_{base}$, $XGLM_{564M}$, $mGPT_{1.3B}$, and $Llama-3.2_{1B}$ are reported in Fig 6.2. Fig 6.2 shows that the BDI performance averaged over 14 BDI directions starts to saturate at 5-shot prompting.

Is few-shot prompting a feasible alternative to supervised fine-tuning methods?: In $mT5_{small}$, 5-shot prompting outperformed supervised fine-tuning except for En-Mn, Mn-En, and En-It. 5-shot prompting outperformed supervised fine-tuning in most BDI directions except En-Mn and Mn-En for $mT5_{base}$ LLM model. In $XGLM_{564M}$, 5-shot prompting outperformed supervised fine-tuning in all the BDI directions. Similarly, 5-shot prompting outperformed supervised fine-tuning in all the BDI directions for $mGPT_{1.3B}$. Evaluation in $Llama-3.2_{1B}$ shows a slightly different result with supervised fine-tuning outperforming 5-shot prompting in En-Mn, En-Fi, En-Tr, En-Ja, Ja-En, and En-Ta. The under-performance of 5-shot prompting in $Llama-3.2_{1B}$ at the above-mentioned BDI directions might be due to many factors like fewer resources of Mn, Fi, Tr, Ja, and Ta languages in the continued training process of $Llama-3.2_{1B}$. The biased nature of the Llama model to resource-rich Languages like English is also a contributing factor. Another factor may be complex morphological understanding in source-to-target BDI tasks like En-Mn, En-Fi, En-Tr, and En-Ta. On the other hand, target-to-source BDI directions like Mn-En, Fi-En, Tr-En, and Ta-Mn are much simpler tasks and thus give better results than supervised fine-tuning by mitigating over-fitting on the training dictionary. While supervised fine-tuning methods are effective for morphologically complex languages, few-shot prompting strikes a balance by leveraging inherent multi-lingual contextual understanding and avoiding over-fitting issues. Few-shot prompting is preferable for practical applications like BDI, where annotated training data is expensive.

6.7 Summary

This chapter evaluates the performance of various large language models (LLMs) across diverse distant language pairs and task settings (unsupervised, supervised fine-tuning, zero-shot, and 5-shot prompting) in intrinsic BDI tasks. The results show that all LLMs consistently perform better on linguistically similar pairs (e.g., En-It) than distant and morphologically complex pairs (e.g., En-Mn, En-Fi, En-Ta). Morphologically rich languages

like Manipuri, Finnish, Turkish, and Tamil pose significant challenges, particularly in unsupervised and zero-shot settings. The 5-shot prompting approach outperforms unsupervised and zero-shot settings in all cases and even surpasses supervised settings in 82.86% of the cases. Few-shot prompting demonstrates robustness against over-fitting, leveraging LLMs' In-context learning multilingual capabilities, making it particularly effective in target-to-source translation even for morphologically complex language pairs. At the same time, few-shot prompting in LLM models like Llama is still ineffective for morphologically rich language pairs like En-Mn and En-Ta in source-to-target BDI tasks. While supervised fine-tuning methods are effective, especially for morphologically complex languages, few-shot prompting strikes a balance by leveraging inherent multi-lingual contextual understanding and avoiding over-fitting issues. Few-shot prompting is preferable for practical applications like BDI, where annotated training data is scarce or expensive. From the analysis, BDI performance saturates at 5-shot prompting, indicating diminishing returns beyond this point. The experimental findings suggest few-shot prompting as a cost-effective and powerful alternative for BDI tasks, with future work focusing on prompt optimization.

The results shows that all LLMs consistently perform better on linguistically similar pairs (e.g., En-It) compared to diverse and morphologically complex pairs. Morphologically rich languages like Manipuri, Finnish, Turkish, and Tamil pose significant challenges, particularly in unsupervised and zero-shot settings. The 5-shot prompting approach outperforms unsupervised and zero-shot settings in all cases and even surpasses supervised settings in 82.86% of cases. Few-shot prompting demonstrates robustness against over-fitting, leveraging LLMs' pretrained multilingual alignment, making it particularly effective for low-resource languages. Few-shot prompting is a powerful approach for leveraging LLMs in linguistically diverse language pairs. While supervised methods are effective, especially for morphologically complex languages, few-shot prompting strikes a balance by maintaining generalization and leveraging contextual embeddings. This makes it a preferable choice for practical applications like BDI where annotated training data is scarce or expensive to obtain.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis investigated the performance of cross-lingual embeddings (CLWEs) for linguistically diverse language pairs on state-of-the-art supervised, weakly supervised, and unsupervised, focusing on English-Manipuri. Key findings include:

1. **Linguistic Challenges**: Significant performance differences were observed between similar language pairs (e.g., English-Italian) and diverse pairs (e.g., English-Manipuri). Linguistic differences and the sparse nature of Manipuri data, resulting from its complex morphological structure, exacerbate these challenges. Also, limited data availability complicates cross-lingual alignment, making bilingual dictionary induction (BDI) challenging.

2. **Improved Dictionary Pair Selection**: The experimental observation shows that diverse language pairs are likely to have poor dictionary pairs, which does not help much in cross-lingual alignment. It is also clear that instead of selecting dictionary pairs based only on the frequency of the source language, it is better to choose dictionary pairs with comparable centrality measures. The centrality-based dictionary pair selection improved alignment quality by emphasizing pairs with semantically similar neighbors and outperformed frequency-based selection methods, enhancing BDI, sentence retrieval, and machine translation tasks.

3. **Proposed Contrastive Learning Method**: The proposed contrastive learning method effectively grouped morphologically related words while separating unrelated ones. It outperformed baseline models like fastText, VecMap, and word translation

contrastive approach, showing improved lexical alignment in BDI, machine translation, and sentence retrieval tasks. The ablation study reveals that the proposed contrastive learning further improves the lexical properties of mBERT/IndicBERT WEs.

4. **Effectiveness of Few-shot Prompting**: Few-shot prompting consistently outperformed unsupervised and zero-shot methods, surpassing supervised methods in 82.86% of cases. In target-to-source BDI, few-shot prompting demonstrates robustness against over-fitting, leveraging LLMs' pre-trained multilingual alignment, making it particularly effective for low-resource and linguistically diverse language pairs. This makes it preferable for practical applications like BDI, where annotated training data is scarce or expensive. The findings suggest few-shot prompting as a cost-effective and powerful alternative for multilingual BDI, with future work focusing on prompt optimization. More studies and large amounts of data are required for morphologically rich languages like Manipuri and Tamil. En-Mn and En-Ta still pose significant challenges, particularly in unsupervised and few-shot settings for source-to-target BDI.

7.2 Limitations and Future Works

This section discusses the limitations associated with the current study and some potential directions to explore in the future. A few of the limitation and major research directions for future explorations of the thesis work are as follows:

- **Poor performance in unsupervised method with regularization**: Evaluation in an unsupervised method like unsupervised VECMAP[5] [42] with regularization using centrality measures fail to improve the performance on BDI task. Instead, it decreases performance, possibly due to the poor quality of the initial seed dictionary generated using the unsupervised approach. Only Degree centrality and Eigenvector centrality are considered. Another more suitable centrality measure is yet to be explored. Besides the dictionary selection task, language-specific linguistics features like morphology and sentence structure can be exploited for better performance in downstream NLP tasks. Evaluation on other extrinsic downstream NLP tasks will also give a better picture of the quality of the cross-lingual embeddings with regularization.

- **mBERT and IndicBert giving better performance than mT5 in MACE framework that uses contrastive learning**: Experiment results under MACE framework that uses contrastive learning show that the proposed method performs

[5] https://github.com/artetxem/vecmap

better in mBERT/IndicBert than mT5. The $loss_{wt}$ gives little performance increase in mT5 compared to mBERT/IndicBert due to overfitting. Evaluation in mT5 with different learning rate as a future direction might shed light on mT5 model under MACE framework.

- **Less data in continue training process of LLM models**: Few-shot prompting in LLM models like Llama is still ineffective for morphologically rich language pairs like En-Mn and En-Ta in source-to-target BDI tasks. This is likely due to model bias nature towards English language and lesser Manipuri and Tamil data in continue training process of Llama model. The major experimental findings suggest few-shot prompting as a cost-effective and powerful alternative for BDI tasks, with future direction on prompt optimization.

Appendix A

A.1 Procrustes problem

The optimization problem is given as:

$$\arg\min_{W} \sum_{i} \|X_{i*}W - Z_{i*}\|^2$$

Expanding the squared norm:

$$\sum_{i} \left(\|X_{i*}W\|^2 + \|Z_{i*}\|^2 - 2X_{i*}WZ_{i*}^T \right)$$

Since $\|Z_{i*}\|^2$ is independent of W, we can ignore it:

$$\arg\min_{W} \sum_{i} \|X_{i*}W\|^2 - 2\sum_{i} X_{i*}WZ_{i*}^T$$

Rewriting in matrix form:

$$\arg\min_{W} \mathrm{Tr}(XWW^TX^T) - 2\mathrm{Tr}(Z^TXW)$$

Since the first term is a quadratic term in W, it can be rewritten using the Frobenius norm:

$$\arg\min_{W} \|XW\|_F^2 - 2\mathrm{Tr}(Z^TXW)$$

In the above expression, Tr( ) denotes the trace operator, which sums all elements along the main diagonal. The last equality follows from the cyclic property of the trace. At this point, we take the Singular Value Decomposition (SVD) of $Z^T X$, given by:

$$Z^T X = U \Sigma V^T$$

Substituting into the trace expression:

$$\mathrm{Tr}(Z^T X W) = \mathrm{Tr}(U \Sigma V^T W) = \mathrm{Tr}(\Sigma V^T W U)$$

Since V^T, W, and U are orthogonal matrices, their product $V^T W U$ remains orthogonal. Additionally, because Σ is a diagonal matrix, its trace is maximized when its diagonal elements remain unchanged after the transformation. This occurs when the orthogonal transformation matrix is the identity, i.e.,

$$V^T W U = I$$

Solving for WW, we obtain the optimal solution:

$$W^* = V U^T$$

Bibliography

[1] Y. Li, F. Liu, N. Collier, A. Korhonen, and I. Vulić, "Improving word translation via two-stage contrastive learning," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 4353–4374. [Online]. Available: https://aclanthology.org/2022.acl-long.299

[2] G. Dinu, A. Lazaridou, and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," *arXiv preprint arXiv:1412.6568*, 2014.

[3] M. Artetxe, G. Labaka, and E. Agirre, "Learning principled bilingual mappings of word embeddings while preserving monolingual invariance," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2289–2294. [Online]. Available: https://www.aclweb.org/anthology/D16-1250

[4] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

[5] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[6] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[7] S. Ruder, I. Vulić, and A. Søgaard, "A survey of cross-lingual word embedding models," *J. Artif. Int. Res.*, vol. 65, no. 1, p. 569–630, May 2019. [Online]. Available: https://doi.org/10.1613/jair.1.11640

[8] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *arXiv preprint arXiv:1309.4168*, 2013.

[9] L. Laitonjam and S. Ranbir Singh, "Manipuri-English machine translation using comparable corpus," in *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*. Virtual: Association for Machine Translation in the Americas, Aug. 2021, pp. 78–88. [Online]. Available: https://aclanthology.org/2021.mtsummit-loresmt.8

[10] L. Laitonjam and S. R. Singh, "Manipuri–english comparable corpus for cross-lingual studies," *Lang. Resour. Eval.*, vol. 57, no. 1, p. 377–413, Feb. 2022. [Online]. Available: https://doi.org/10.1007/s10579-021-09576-y

[11] A. Søgaard, S. Ruder, and I. Vulić, "On the limitations of unsupervised bilingual dictionary induction," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 778–788. [Online]. Available: https://aclanthology.org/P18-1072

[12] M. Zhang, K. Xu, K.-i. Kawarabayashi, S. Jegelka, and J. Boyd-Graber, "Are girls neko or shōjo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3180–3189. [Online]. Available: https://aclanthology.org/P19-1307

[13] B. Patra, J. R. A. Moniz, S. Garg, M. R. Gormley, and G. Neubig, "Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 184–193. [Online]. Available: https://aclanthology.org/P19-1018

[14] M. Zhang, Y. Liu, H. Luan, and M. Sun, "Earth mover's distance minimization for unsupervised bilingual lexicon induction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1934–1945. [Online]. Available: https://aclanthology.org/D17-1207

[15] L. Laitonjam and S. R. Singh, "Manipuri–english comparable corpus for cross-lingual studies," *Language Resources and Evaluation*, pp. 1–37, 2022.

[16] T. D. Singh and S. Bandyopadhyay, "Statistical machine translation of English-Manipuri using morpho-syntactic and semantic information," in *Proceedings of the 9th Conference of the Association for Machine Translation in the*

Americas: Student Research Workshop. Denver, Colorado, USA: Association for Machine Translation in the Americas, Oct. 31-Nov. 4 2010. [Online]. Available: https://aclanthology.org/2010.amta-srw.1

[17] S. I. Choudhury, L. S. Singh, S. Borgohain, and P. K. Das, "Morphological analyzer for manipuri: design and implementation," in *Asian Applied Computing Conference*. Springer, 2004, pp. 123–129.

[18] Y. Li, A. Korhonen, and I. Vulić, "On bilingual lexicon induction with large language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 9577–9599. [Online]. Available: https://aclanthology.org/2023.emnlp-main.595

[19] Y. Liu, X. Chen, G. Xing, J. Zhang, and R. Yan, "IAD: In-context learning ability decoupler of large language models in meta-training," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 8535–8545. [Online]. Available: https://aclanthology.org/2024.lrec-main.749

[20] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 462–471.

[21] S. Gouws and A. Søgaard, "Simple task-specific bilingual word embeddings," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May–Jun. 2015, pp. 1386–1390. [Online]. Available: https://www.aclweb.org/anthology/N15-1157

[22] L. Duong, H. Kanayama, T. Ma, S. Bird, and T. Cohn, "Learning crosslingual word embeddings without bilingual corpora," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1285–1295. [Online]. Available: https://www.aclweb.org/anthology/D16-1136

[23] M. Artetxe, G. Labaka, and E. Agirre, "Learning bilingual word embeddings with (almost) no bilingual data," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 451–462. [Online]. Available: https://www.aclweb.org/anthology/P17-1042

[24] Y. Doval, J. Camacho-Collados, L. Espinosa-Anke, and S. Schockaert, "Improving cross-lingual word embeddings by meeting in the middle," *arXiv preprint arXiv:1808.08780*, 2018.

[25] M. Biesialska and M. R. Costa-jussà, "Refinement of unsupervised cross-lingual word embeddings," in *European Conference on Artificial Intelligence*. Santiago, Galicia: Frontiers in Artificial Intelligence and Applications, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211252854

[26] C. Xing, D. Wang, C. Liu, and Y. Lin, "Normalized word embedding and orthogonal transform for bilingual word translation," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, May–Jun. 2015, pp. 1006–1011. [Online]. Available: https://www.aclweb.org/anthology/N15-1104

[27] A. Üstün, G. Bouma, and G. van Noord, "Cross-lingual word embeddings for morphologically rich languages," in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria: INCOMA Ltd., Sep. 2019, pp. 1222–1228. [Online]. Available: https://aclanthology.org/R19-1140

[28] S. Chimalamarri, D. Sitaram, and A. Jain, "Morphological segmentation to improve crosslingual word embeddings for low resource languages," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 19, no. 5, jun 2020. [Online]. Available: https://doi.org/10.1145/3390298

[29] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[30] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[31] A. Lazaridou, G. Dinu, and M. Baroni, "Hubness and pollution: Delving into cross-space mapping for zero-shot learning," in *Proceedings of the 53rd Annual*

Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 270–280. [Online]. Available: https://www.aclweb.org/anthology/P15-1027

[32] M. Xiao and Y. Guo, "Distributed word representation learning for cross-lingual dependency parsing," in *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan: Association for Computational Linguistics, Jun. 2014, pp. 119–129. [Online]. Available: https://www.aclweb.org/anthology/W14-1613

[33] D. Naorem, O. J. Singh, S. R. Singh, and P. Sarmah, "English-manipuri cross-lingual embedding: A preliminary study," in *2023 International Conference on Asian Language Processing (IALP)*, 2023, pp. 74–79.

[34] D. Naorem, S. R. Singh, and P. Sarmah, "Embarking on a preliminary exploration: Cross-lingual embedding in english-manipuri," *International Journal of Asian Language Processing*.

[35] S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," *arXiv preprint arXiv:1702.03859*, 2017.

[36] A. Conneau, G. Lample, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," *arXiv preprint arXiv:1710.04087*, 2017.

[37] M. Zhang, Y. Liu, H. Luan, and M. Sun, "Adversarial training for unsupervised bilingual lexicon induction," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1959–1970. [Online]. Available: https://www.aclweb.org/anthology/P17-1179

[38] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 2672–2680.

[39] Z. Mhammedi, A. Hellicar, A. Rahman, and J. Bailey, "Efficient orthogonal parametrisation of recurrent neural networks using householder reflections," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2401–2409.

[40] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[41] A. Søgaard, S. Ruder, and I. Vulić, "On the limitations of unsupervised bilingual dictionary induction," *arXiv preprint arXiv:1805.03620*, 2018.

[42] M. Artetxe, G. Labaka, and E. Agirre, "A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 789–798. [Online]. Available: https://www.aclweb.org/anthology/P18-1073

[43] M. Biesialska and M. R. Costa-jussà, "Refinement of unsupervised cross-lingual word embeddings," *arXiv preprint arXiv:2002.09213*, 2020.

[44] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[45] I. Vulić, G. Glavaš, R. Reichart, and A. Korhonen, "Do we really need fully unsupervised cross-lingual embeddings?" *arXiv preprint arXiv:1909.01638*, 2019.

[46] M. Zhang, K. Xu, K.-i. Kawarabayashi, S. Jegelka, and J. Boyd-Graber, "Are girls neko or sh\= ojo? cross-lingual alignment of non-isomorphic embeddings with iterative normalization," *arXiv preprint arXiv:1906.01622*, 2019.

[47] P. Czarnowska, S. Ruder, E. Grave, R. Cotterell, and A. Copestake, "Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 974–983. [Online]. Available: https://aclanthology.org/D19-1090

[48] J. Khatri, R. Murthy, and P. Bhattacharyya, "A study of efficacy of cross-lingual word embeddings for indian languages," in *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, ser. CoDS COMAD 2020. New York, NY, USA: Association for Computing Machinery, 2020, p. 347–348. [Online]. Available: https://doi.org/10.1145/3371158.3371219

[49] T. D. Singh and S. Bandyopadhyay, "Manipuri-English bidirectional statistical machine translation systems using morphology and dependency relations," in *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*. Beijing, China: Coling 2010 Organizing Committee, Aug. 2010, pp. 83–91. [Online]. Available: https://aclanthology.org/W10-3811

[50] L. Laitonjam and S. R. Singh, "Manipuri-english cross-lingual word embeddings using a temporally aligned comparable corpus," in *2021 International Conference on Asian Language Processing (IALP)*, 2021, pp. 195–199.

[51] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[52] A. Kunchukuttan, M. Khapra, G. Singh, and P. Bhattacharyya, "Leveraging orthographic similarity for multilingual neural transliteration," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 303–316, 2018.

[53] L. Laitonjam, L. G. Singh, and S. R. Singh, "Transliteration of english loanwords and named-entities to manipuri: Phoneme vs grapheme representation," in *2018 International Conference on Asian Language Processing (IALP)*. IEEE, 2018, pp. 255–260.

[54] P. Koehn, "Europarl: A parallel corpus for statistical machine translation," in *Proceedings of machine translation summit x: papers*, 2005, pp. 79–86.

[55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[56] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[57] S. Doddapaneni, R. Aralikatte, G. Ramesh, S. Goyal, M. M. Khapra, A. Kunchukuttan, and P. Kumar, "Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 12 402–12 426. [Online]. Available: https://aclanthology.org/2023.acl-long.693

[58] J. H. Paik, M. Mitra, S. K. Parui, and K. Järvelin, "Gras: An effective and efficient stemming algorithm for information retrieval," *ACM Trans. Inf. Syst.*, vol. 29, no. 4, dec 2011. [Online]. Available: https://doi.org/10.1145/2037661.2037664

[59] A. Klementiev, I. Titov, and B. Bhattarai, "Inducing crosslingual distributed representations of words," in *Proceedings of COLING 2012*, 2012, pp. 1459–1474.

[60] I. Vulić and M.-F. Moens, "Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 363–372.

[61] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Sy2ogebAW

[62] C.-H. Lee and H.-Y. Lee, "Cross-lingual transfer learning for question answering," *arXiv preprint arXiv:1907.06042*, 2019.

[63] T. Brychcín, S. Taylor, and L. Svoboda, "Cross-lingual word analogies using linear transformations between semantic spaces," *Expert Systems with Applications*, vol. 135, pp. 287–295, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419304191

[64] N. Taghizadeh and H. Faili, "Cross-lingual transfer learning for relation extraction using universal dependencies," *Computer Speech and Language*, vol. 71, p. 101265, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821000711

[65] Y.-C. Wang, C.-M. Chuang, C.-K. Wu, C.-L. Pan, and R. T.-H. Tsai, "Cross-language article linking with deep neural network based paragraph encoding," *Computer Speech and Language*, vol. 72, p. 101279, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821000826

[66] Z. Ahmad, R. Jindal, A. Ekbal, and P. Bhattachharyya, "Borrow from rich cousin: transfer learning for emotion detection using cross lingual embedding," *Expert Systems with Applications*, vol. 139, p. 112851, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419305536

[67] R. Catelli, L. Bevilacqua, N. Mariniello, V. Scotto di Carlo, M. Magaldi, H. Fujita, G. De Pietro, and M. Esposito, "Cross lingual transfer learning for sentiment analysis of italian tripadvisor reviews," *Expert Systems with Applications*, vol. 209, p. 118246, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422013926

[68] H. Aldarmaki and M. Diab, "Context-aware cross-lingual mapping," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3906–3911. [Online]. Available: https://aclanthology.org/N19-1391

[69] I. Vulić, E. M. Ponti, R. Litschko, G. Glavaš, and A. Korhonen, "Probing pretrained language models for lexical semantics," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, Nov. 2020, pp. 7222–7240. [Online]. Available: https://aclanthology.org/2020.emnlp-main.586

[70] T. Mickus, D. Paperno, M. Constant, and K. van Deemter, "What do you mean, BERT?" in *Proceedings of the Society for Computation in Linguistics 2020*, A. Ettinger, G. Jarosz, and J. Pater, Eds. New York, New York: Association for Computational Linguistics, Jan. 2020, pp. 279–290. [Online]. Available: https://aclanthology.org/2020.scil-1.35

[71] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=H196sainb

[72] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," *CoRR*, vol. abs/1309.4168, 2013. [Online]. Available: http://arxiv.org/abs/1309.4168

[73] Y. Li, K. Yu, and Y. Zhang, "Learning cross-lingual mappings in imperfectly isomorphic embedding spaces," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2630–2642, 2021.

[74] G. Dinu and M. Baroni, "Improving zero-shot learning by mitigating the hubness problem," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6568

[75] S. L. Smith, D. H. P. Turban, S. Hamblin, and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=r1Aab85gg

[76] I. Vulić, G. Glavaš, R. Reichart, and A. Korhonen, "Do we really need fully unsupervised cross-lingual embeddings?" in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 4407–4418. [Online]. Available: https://aclanthology.org/D19-1449

[77] S.-i. Amari, "Backpropagation and stochastic gradient descent method," *Neurocomputing*, vol. 5, no. 4-5, pp. 185–196, 1993.

[78] B. K. P. Horn, H. M. Hilden, and S. Negahdaripour, "Closed-form solution of absolute orientation using orthonormal matrices," *J. Opt. Soc. Am. A*, vol. 5, no. 7, pp. 1127–1135, Jul 1988. [Online]. Available: http://opg.optica.org/josaa/abstract.cfm?URI=josaa-5-7-1127

[79] B. Haddow and F. Kirefu, "PMIndia – A Collection of Parallel Corpora of Languages of India," *arXiv e-prints*, p. arXiv:2001.09907, Jan 2020.

[80] A. Kunchukuttan, P. Mehta, and P. Bhattacharyya, "The IIT Bombay English-Hindi parallel corpus," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1548

[81] R. Pryzant, Y. Chung, D. Jurafsky, and D. Britz, "JESC: Japanese-English subtitle corpus," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1182

[82] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13. Red Hook, NY, USA: Curran Associates Inc., 2013, p. 3111–3119.

[83] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[84] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Hubs in space: Popular nearest neighbors in high-dimensional data," *Journal of Machine Learning Research*, vol. 11, no. sept, pp. 2487–2531, 2010.

[85] I. Vulić, E. M. Ponti, A. Korhonen, and G. Glavaš, "LexFit: Lexical fine-tuning of pretrained language models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Co@articleschonemann1966generalized, title=A generalized solution of the orthogonal procrustes problem, author=Schönemann, Peter H, journal=Psychometrika, volume=31, number=1, pages=1–10, year=1966, publisher=Springer nference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 5269–5283. [Online]. Available: https://aclanthology.org/2021.acl-long.410

[86] I. Vulić, N. Mrkšić, R. Reichart, D. Ó Séaghdha, S. Young, and A. Korhonen, "Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, R. Barzilay and M.-Y. Kan, Eds. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 56–68. [Online]. Available: https://aclanthology.org/P17-1006

[87] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mT5: A massively multilingual pre-trained text-to-text transformer," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 483–498. [Online]. Available: https://aclanthology.org/2021.naacl-main.41

[88] K. Gupta, B. Th'erien, A. Ibrahim, M. L. Richter, Q. G. Anthony, E. Belilovsky, I. Rish, and T. Lesort, "Continual pre-training of large language models: How

to (re)warm your model?" *ArXiv*, vol. abs/2308.04014, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:260704601

[89] D. Naorem, S. R. Singh, and P. Sarmah, "Improving linear orthogonal mapping based cross-lingual representation using ridge regression and graph centrality," *Computer Speech and Language*, vol. 87, p. 101640, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230824000238

[90] P. Czarnowska, S. Ruder, R. Cotterell, and A. Copestake, "Morphologically aware word-level translation," in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 2847–2860. [Online]. Available: https://aclanthology.org/2020.coling-main.256

[91] Z. Mao, C. Chu, and S. Kurohashi, "Ems: Efficient and effective massively multilingual sentence embedding learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2841–2856, 2024.

[92] P. H. Schönemann, "A generalized solution of the orthogonal procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[93] T. Viklands, "Algorithms for the weighted orthogonal procrustes problem and other least squares problems," Ph.D. dissertation, Umeå University, Computing Science, 2006.

[94] M. Bañón, M. Chichirau, M. Esplà-Gomis, M. L. Forcada, A. Galiano-Jiménez, C. García-Romero, T. Kuzman, N. Ljubešić, R. van Noord, L. Pla Sempere, G. Ramírez-Sánchez, P. Rupnik, V. Suchomel, A. Toral, and J. Zaragoza-Bernabeu, "Turkish-english parallel corpus MaCoCu-tr-en 2.0," 2023, slovenian language resource repository CLARIN.SI. [Online]. Available: http://hdl.handle.net/11356/1816

[95] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017. [Online]. Available: https://aclanthology.org/Q17-1010

[96] A. D. et al., "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[97] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, W. Ammar, A. Louis, and N. Mostafazadeh, Eds.

Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 48–53. [Online]. Available: https://aclanthology.org/N19-4009

[98] T. Tang, W. Luo, H. Huang, D. Zhang, X. Wang, X. Zhao, F. Wei, and J.-R. Wen, "Language-specific neurons: The key to multilingual capabilities in large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 5701–5715. [Online]. Available: https://aclanthology.org/2024.acl-long.309/

[99] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, Jan. 2023. [Online]. Available: https://doi.org/10.1145/3560815

[100] Z. Cheng, J. Kasai, and T. Yu, "Batch prompting: Efficient inference with large language model APIs," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, M. Wang and I. Zitouni, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 792–810. [Online]. Available: https://aclanthology.org/2023.emnlp-industry.74/

[101] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel, "ByT5: Towards a token-free future with pre-trained byte-to-byte models," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 291–306, 2022. [Online]. Available: https://aclanthology.org/2022.tacl-1.17

[102] X. V. Lin, T. Mihaylov, M. Artetxe, T. Wang, S. Chen, D. Simig, M. Ott, N. Goyal, S. Bhosale, J. Du, R. Pasunuru, S. Shleifer, P. S. Koura, V. Chaudhary, B. O'Horo, J. Wang, L. Zettlemoyer, Z. Kozareva, M. Diab, V. Stoyanov, and X. Li, "Few-shot learning with multilingual generative language models," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 9019–9052. [Online]. Available: https://aclanthology.org/2022.emnlp-main.616

[103] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," *ArXiv*, vol. abs/2008.00401, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220936592

[104] O. Shliazhko, A. Fenogenova, M. Tikhonova, A. Kozlova, V. Mikhailov, and T. Shavrina, "mgpt: Few-shot learners go multilingual," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 58–79, 2024.

[105] D. Naorem, S. R. Singh, and P. Sarmah, "Embarking on a preliminary exploration: Cross-lingual embedding in english-manipuri," *International Journal of Asian Language Processing*, vol. 34, no. 02, p. 2450007, 2024. [Online]. Available: https://doi.org/10.1142/S2717554524500073

Publications

Conference:

1. D. Naorem, O. J. Singh, S. R. Singh and P. Sarmah, "English-Manipuri Cross-Lingual Embedding: A Preliminary Study," 2023 International Conference on Asian Language Processing (IALP), Singapore, Singapore, 2023, pp. 74-79, doi: 10.1109/IALP61005.2023.10337115.

2. Few-shot Prompting or Fully Supervised? A Comparative Study of LLMs for Linguistically Distant Language Pairs in BDI (**Manuscript under Preparation**)

Journal:

1. Deepen Naorem, Sanasam Ranbir Singh, Priyankoo Sarmah, Improvin linear orthogonal mapping based cross-lingual representation using ridge regression and graph centrality, Computer Speech & Language, Volume 87, 2024, 101640, ISSN 0885-2308, https://doi.org/10.1016/j.csl.2024.101640.

2. Naorem, D., Singh, S. R., Sarmah, P. (2024). Embarking on a Preliminary Exploration: Cross-Lingual Embedding in English-Manipuri. International Journal of Asian Language Processing. 10.1142/s2717554524500073.

3. D. Naorem, S. R. Singh, T. J. Singh and P. Sarmah, "MACE: Morphology Aware Cross-Lingual Embedding Using Contrastive Learning," in IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 3124-3136, 2025, doi: 10.1109/TASLPRO.2025.3589860.

Brief Biography of the Author

Deepen Naorem was born on September 15, 1989, in Imphal-West, Manipur, India. He earned a Bachelor of Engineering (B.E.) degree in Computer Science and Engineering from Anna University, Chennai, in 2012. In 2013, he qualified for the M.Tech program through the PGCET conducted by the Karnataka Examination Authority and subsequently completed his M.Tech at CMRIT, Bangalore, in September 2016. Afterward, he worked as a Junior Research Fellow (JRF) in the Department of Computer Science at Manipur University from January 2017 to July 2017. He then served as a JRF in the Department of Computer Science and Engineering at IIT Guwahati from July 2017 to December 2018. Deepen later enrolled as a Ph.D. research scholar at the Center for Linguistic Science and Technology, Indian Institute of Technology (IIT) Guwahati, under the joint supervision of Prof. Sanasam Ranbir Singh and Prof. Priyankoo Sarmah. His research interests include Cross-lingual Embeddings, Multilingual Representation, Natural Language Processing (NLP), Sentiment Analysis, Machine Learning, and Deep Learning.