

BhashaSetu: Cross-Lingual Knowledge Transfer from High-Resource to Extreme Low-Resource Languages

Subhadip Maji Arnab Bhattacharya

Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur
{subhadip,arnabb}@cse.iitk.ac.in

Abstract

Despite remarkable advances in natural language processing, developing effective systems for low-resource languages remains a formidable challenge, with performances typically lagging far behind high-resource counterparts due to data scarcity and insufficient linguistic resources. Cross-lingual knowledge transfer has emerged as a promising approach to address this challenge by leveraging resources from high-resource languages. In this paper, we investigate methods for transferring linguistic knowledge from high-resource languages to low-resource languages, where the number of labeled training instances is in hundreds. We focus on sentence-level and word-level tasks. We introduce a novel method, GETR (Graph-Enhanced Token Representation) for cross-lingual knowledge transfer along with two adopted baselines (a) augmentation in hidden layers and (b) token embedding transfer through token translation. Experimental results demonstrate that our GNN-based approach significantly outperforms existing multilingual and cross-lingual baseline methods, achieving 13 percentage point improvements on truly low-resource languages (Mizo, Khasi) for POS tagging, and 20 and 27 percentage point improvements in macro-F1 on simulated low-resource languages (Marathi, Bangla, Malayalam) across sentiment classification and NER tasks respectively. We also present a detailed analysis of the transfer mechanisms and identify key factors that contribute to successful knowledge transfer in this linguistic context.

1 Introduction

Cross-lingual knowledge transfer has emerged as a crucial approach for improving natural language processing capabilities across different languages. Recent advances in multilingual model variants have demonstrated remarkable success in this domain by jointly training on multiple languages simultaneously, enabling zero-shot and few-shot

learning capabilities. These models, such as XLM-R (Conneau et al., 2020), XLM-V (Liang et al., 2023) and mmBERT (Marone et al., 2025), learn shared representations across languages, thereby facilitating knowledge transfer from high-resource to low-resource languages. The success of these models largely stems from their ability to leverage massive multilingual corpora and transformer-based architectures (Vaswani et al., 2017), which effectively capture cross-lingual patterns and relationships.

However, when dealing with extremely low-resource scenarios where target languages have very limited labeled data (e.g., only 100 training instances), even state-of-the-art multilingual models struggle (Wu and Dredze, 2020; Downey et al., 2024; Cassano et al., 2024) to generalize effectively. While parameter-efficient techniques like Adapter fine-tuning (Houlsby et al., 2019) extreme and LoRA (Hu et al., 2021) reduce overfitting by updating fewer parameters, they still underperform in such extreme low-resource settings. Cross-lingual models like AdaMergeX (Zhao et al., 2025) also fail to capture sufficient linguistic nuances when number of target language examples is low.

Most languages worldwide have extremely limited digital resources – India alone has at least hundreds of such languages, including several of its 22 official languages like Dogri, Bodo, Kashmiri and Santali. For practical evaluation in this work, we use comparatively higher-resource languages (Marathi, Bangla, Malayalam) as *simulated* low-resource scenarios to enable rigorous testing with sufficient evaluation data. To validate our approach on *truly* low-resource languages, we also evaluate on two real low-resource languages: Mizo and Khasi, which have only 502 and 507 annotated sentences respectively for POS tagging tasks.

To address this challenge of extreme low-resource setting, we propose a comprehensive framework that intelligently transfers linguistic

knowledge from high-resource to low-resource languages through one novel and two adopted complementary approaches. We name our approach **BhashaSetu** after the words “Bhasha” and “Setu” that mean “language” and “bridge” respectively in most Indian languages, highlighting its role in bridging languages.

Our approach is as follows. First, we introduce an adopted baseline, Hidden Augmentation Layers (HAL) that create mixed representations in the hidden space, allowing controlled knowledge transfer while preserving the target language’s distinctive features. Second, we develop another adopted baseline, a token embedding transfer mechanism that leverages translation-based mappings to initialize low-resource language embeddings effectively. Finally, we propose a novel Graph-Enhanced Token Representation (GETR) approach that uses Graph Neural Networks (Zhou et al., 2020; Kipf and Welling, 2016; Veličković et al., 2017) to enable dynamic knowledge sharing between languages at the token level, thereby capturing complex cross-lingual relationships through graph-based message passing. In short, our contributions are:

1. We propose a comprehensive framework, **BHASHASETU**, for cross-lingual knowledge transfer in extreme low-resource scenarios, comprising two adopted baselines such as hidden augmentation layer (HAL) and token embedding transfer (TET), and a novel graph-enhanced token representation (GETR) with GNNs (Sec. 3).
2. We conduct extensive experiments across multiple NLP tasks and language pairs spanning multiple languages, demonstrating the versatility and robustness of our approach. Experimental results on POS tagging task low-resource languages Mizo and Khasi using high-resource languages Hindi and English demonstrate that our novel GNN-based approach significantly outperforms existing methods, achieving 13 percentage points improvement respectively in macro-F1 score compared to traditional multilingual and cross-lingual baselines while requiring only 100 training instances in the low-resource language (Sec. 4).
3. We provide systematic analysis of the impact of various factors on cross-lingual knowledge transfer, including mixing coefficient, architectural depth and dataset size ratios between languages (Sec. 4).

2 Related Work

Cross-lingual transfer learning has advanced significantly with multilingual pre-trained models such as XLM-R (Conneau et al., 2020) and mBERT (Marone et al., 2025). While effective, these approaches require substantial multilingual training data, limiting their applicability in extreme low-resource settings. Recent parameter-efficient fine-tuning methods like LoRA (Hu et al., 2021), AdaMergeX (Zhao et al., 2025), and SALT (Lee et al., 2025) reduce overfitting risks but struggle with extremely limited target data.

Data augmentation techniques in hidden spaces, including mixup (Zhang et al., 2017; Verma et al., 2018) and their NLP adaptations (Chen et al., 2020; Sun et al., 2020), have proven valuable for low-resource scenarios and are comprehensively surveyed by Feng et al. (2021). Token-level transfer approaches like trans-tokenization (Remy et al., 2024) enable cross-lingual embedding transfer without requiring parallel data, addressing a critical challenge for low-resource languages.

Graph-based cross-lingual methods such as Heterogeneous GNNs (Wang et al., 2021) depend on external semantic parsers and operate solely at the GNN level, without integrating graph knowledge into transformer models. Colexification-based multilingual graphs (Liu et al., 2023) construct graphs from colexification relations rather than token interactions, and similarly do not infuse graph information into transformers. While recent work has employed graph-based transformers with UCCA semantic graphs (Nguyen et al., 2023), such approaches require pre-trained semantic parsers that are typically unavailable for low-resource Indian languages. In contrast, our GETR method constructs token-level graphs directly from training data and uniquely integrates GNN-based token interactions within the transformer, enabling dynamic, fine-grained cross-lingual knowledge sharing without external linguistic resources.

3 Methodology

This section presents three approaches for cross-lingual knowledge transfer: (a) **Augmentation in Hidden Layers (HAL)**, an adaptation of existing layer-wise mixing strategies; (b) **Token Embedding Transfer through Translation (TET)**, a refinement of token embedding transfer through translation dictionaries; and (c) **Graph-Enhanced Token Representation (GETR)**, a novel approach

leveraging Graph Neural Networks to dynamically share token embeddings across languages at hidden layers. While HAL and TET build upon established techniques from prior work, both are adapted specifically for the cross-lingual setting with modifications tailored to leverage high-resource language knowledge for low-resource language tasks. In contrast, GETR introduces a fundamentally new mechanism that constructs dynamic token-level graphs to enable fine-grained cross-lingual knowledge sharing through neighborhood aggregation in GNNS. Before delving into the technical details of these approaches, we first formally define the problem statement.

Problem Statement: Let us formally define our notation for cross-lingual knowledge transfer. For a high-resource language, we denote the dataset of textual instances as $\mathbf{X}_H = \{x_1, x_2, \dots, x_{N_H}\}$, where each x_i represents an individual text instance (e.g., a sentence). The corresponding task-specific outputs are represented as $Y_H = \{y_1, y_2, \dots, y_{N_H}\}$. Here, N_H represents the total number of instances in the high-resource dataset, typically in the order of thousands or more. Similarly, we denote the low-resource language dataset as \mathbf{X}_L and its corresponding outputs as Y_L , where $|\mathbf{X}_L| = N_L \ll N_H$, with N_L being extremely small (approximately 100 instances). This extreme data scarcity in the low-resource setting presents the core challenge in our task.

We define the combined dataset as $\mathbf{X} = \{\mathbf{X}_H \cup \mathbf{X}_L\}$ and $Y = \{Y_H \cup Y_L\}$. Our objective is to learn a model $\mathbf{M} : \mathbf{X} \rightarrow Y$ that maps input text instances from either or both \mathbf{X}_H and \mathbf{X}_L to their respective outputs, while effectively leveraging the high-resource language data to compensate for the limited low-resource samples. The output space Y can correspond to any encoder-based task, with two common task variants. The first is for sentence-level tasks (such as sentiment analysis) where y_i is one of c classes. The second is for sequence-labeling tasks (such as NER): $y_i = [y_{i1}, y_{i2}, \dots, y_{iT}]$, where T is the sequence length and each token-level label $y_{it} \in \mathcal{Y}_{tags}$ represents a class (such as an NER tag).

Despite the different output structures, the core challenge of effective cross-lingual knowledge transfer remains consistent across tasks, allowing us to apply the same methodological approaches with task-specific adaptations. We next describe the three methods.

3.1 HAL

Hidden layer augmentation has emerged as a prevalent technique for generating synthetic training data in the latent space when working with textual inputs (Zhang et al., 2017; Verma et al., 2018). While this approach has been successfully applied for domain adaptation within the same language (Zhang et al., 2022), its application to cross-lingual knowledge transfer, particularly from high-resource to low-resource languages, represents a novel direction. This method is particularly versatile as it can be applied to any high-resource and low-resource language pair, regardless of their script similarities.

Let $\mathbf{E}_M : \mathbf{X} \rightarrow \mathbf{H}$ denote the encoder component of the model \mathbf{M} that maps each input text x_i to its final encoded CLS representation h_i^{CLS} . We propose a hidden augmentation mechanism that fuses knowledge from high-resource and low-resource languages through a weighted combination in the latent space. Formally, we generate new training pairs $A_i = (h_{A_i}^{\text{CLS}}, y_{A_i})$ as follows:

$$h_{A_i}^{\text{CLS}} = \alpha \cdot h_{H_i}^{\text{CLS}} + (1 - \alpha) \cdot h_{L_i}^{\text{CLS}} \quad (1)$$

$$y_{A_i} = \begin{cases} \alpha \cdot y_{H_i} + (1 - \alpha) \cdot y_{L_i}, \\ (\alpha \cdot y_{H_i,t} + (1 - \alpha) \cdot y_{L_i,t})_{t=1}^T \end{cases} \quad (2)$$

where $\alpha \in [0, 1]$ is a mixing coefficient that controls the contribution of each language. This coefficient can be either fixed through training or randomly sampled per iteration. For sentence tasks with c classes, $y_{H_i}, y_{L_i} \in \mathbb{R}^c$ are typically one-hot encoded vectors, while for sequence tasks, $y_{H_i,t}, y_{L_i,t} \in \mathbb{R}^{|\mathcal{Y}_{tags}|}$ represent the tag distribution at position t .

Empirically, α values between 0.1 and 0.4 yield optimal results, as they maintain the primary characteristics of the low-resource language while supplementing it with knowledge from the high-resource language. Since the augmentation produces soft labels, we employ KL-divergence loss (Cui et al., 2024) instead of standard cross-entropy loss (Mao et al., 2023) for soft labels and cross-entropy for hard labels during training. This framework can be further extended by adding multiple transformer layers above E_M and performing augmentation at each layer’s CLS output, thus enabling hierarchical knowledge fusion.

3.2 TET

Traditional approaches often initialize token embeddings for low-resource languages randomly,

Algorithm 1 Token Embedding Transfer through Translation (TET)

```

1:  $V_L \leftarrow$  Set of unique words from LRL corpus
2: for all  $w_l \in V_L$  do ▷ For each LRL word
3:    $w_h \leftarrow \text{TranslateToHRL}(w_l)$ 
4:    $T_h \leftarrow \text{HRLTokenize}(w_h)$ 
5:    $T_l \leftarrow \text{LRLTokenize}(w_l)$ 
6:    $E_h \leftarrow \{\text{GetPretrainedEmbeddings}(t) | t \in T_h\}$  ▷
   HRL token embeddings
7:    $e_{\text{avg}} \leftarrow \text{Mean}(E_h)$ 
8:   for all  $t_l \in T_l$  do ▷ For each LRL token
9:      $P_{t_l} \leftarrow \emptyset$  ▷ Initialize projected embeddings set
10:    for all  $w' \in V_L$  do ▷ Check all LRL words
11:      if  $t_l \in \text{LRLTokenize}(w')$  then
12:         $P_{t_l} \leftarrow P_{t_l} \cup \{e_{\text{avg}}\}$ 
13:      end if
14:    end for
15:     $E_l[t_l] \leftarrow \text{Mean}(P_{t_l})$  ▷ Final embedding for LRL
   token
16:   end for
17: end for
18: return  $E_l$  ▷ Dictionary of LRL token embeddings

```

which can lead to suboptimal performance, especially when training data is scarce. We propose an initialization strategy that leverages token embeddings from a high-resource language through translation mapping (Remy et al., 2024). This approach provides a more informed starting point for the embedding matrix of the low-resource language, thereby enabling effective fine-tuning even with limited training samples. The core idea is to initialize the token embeddings of the low-resource language using the semantic information captured in the pre-trained embeddings of their translated counterparts in the high-resource language. While this method assumes the availability of word-level translations for the training data of the low-resource language, it does not require any pre-trained models or large corpora in the low-resource language. In our experiments, we used pymultidictionary (Pizarro, 2025) primarily to make our translation process seamless, faster and automated for the languages in our study, followed by manual verification of the translations to ensure accuracy. For extremely low-resource languages without dictionary support, we recommend manually translating the limited training vocabulary (which is manageable given the small dataset size of ~ 100 instances).

Algorithm 1 details our systematic process for transferring token embeddings from a high-resource language (e.g., English) to a low-resource language (e.g., Marathi). To illustrate this process, consider transferring embeddings for the Marathi word "āntarbhāṣika" meaning "cross-lingual" in English. The word would be translated to En-

glish as "cross-lingual", which might be tokenized as "cross" + "lingual" in English. The word, "āntarbhāṣika" would be tokenized in Marathi, potentially splitting it into subword tokens like "āntar" + "bhāṣika". The pre-trained embeddings for these English tokens are retrieved and averaged. For each Marathi token, we collect all instances where it appears across different words in the Marathi corpus. For example, the token "bhāṣika" might also appear in words like "bahubhāṣika" (meaning "multilingual"). Finally, we average all corresponding English embedding projections to create the final embedding for each Marathi token. While we show transliterated examples here for clarity, in our actual experiments we used the original scripts for all languages.

3.3 GETR

We propose a novel approach leveraging Graph Neural Networks (GNN) (Zhou et al., 2020) to enable dynamic knowledge sharing between high-resource and low-resource languages at the token level. For each batch of mixed-language inputs, we construct an undirected graph $G = (T, C)$, where $T = \{t_1, t_2, \dots, t_{N_k}\}$ represents the set of N unique tokens in batch k . The edge set C captures sequential relationships between tokens, defined as $C \subseteq \{t_{ij}, t_{i(j+1)} | t_{ij}, t_{i(j+1)} \in T\}$, where tokens $t_{i1}, t_{i2}, \dots, t_{in}$ form sentence s_i .

To illustrate the mechanism, consider two sentences: "The movie was good" from a high-resource language and "I was impressed with the movie" from a low-resource language. As shown in Figure 1, tokens are represented as nodes with edges connecting consecutive tokens within each sentence. When computing the representation for shared tokens (e.g., "was"), the model incorporates contextual information from both language environments. This allows the CLS embedding of the low-resource sentence to benefit from the high-resource language's token representations through neighborhood aggregation.

Given the encoder output $\mathbf{H} \in \mathbb{R}^{B \times S \times D}$ (where B , S , and D denote batch size, sequence length, and embedding dimension respectively), we reshape it to $\mathbf{H}' \in \mathbb{R}^{L \times D}$ ($L = B \times S$) for GNN processing. We employ either GCN (Kipf and Welling, 2016) or GAT (Veličković et al., 2017) layers with an adjacency matrix $\mathbf{A} \in \{0, 1\}^{L \times L}$ that captures token relationships such as $\mathbf{A}_{ij} = 1$ if l_i and l_j are consecutive tokens in a sentence. Notably, we construct \mathbf{A} using the flattened dimension L rather

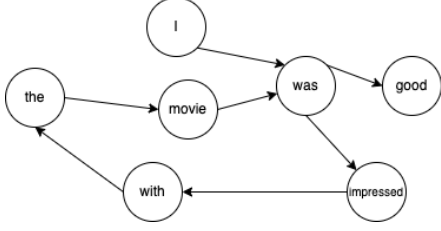


Figure 1: Graphical representation of tokens of two sentences in a batch: “The movie was good” and “I was impressed with the movie”.

than unique tokens, allowing for token repetition which makes the array multiplication simpler and straight-forward. The GNN output is then reshaped to generate query \mathbf{Q} and key \mathbf{K} matrices for the subsequent transformer layer:

$$\begin{aligned} \mathbf{H}' &= \text{Reshape}(\mathbf{H}) \in \mathbb{R}^{L \times D} \\ \mathbf{H}'_{\mathbf{G}} &= \text{GNN}(\mathbf{H}') \\ \mathbf{H}_{\mathbf{G}} &= \text{Reshape}(\mathbf{H}'_{\mathbf{G}}) \in \mathbb{R}^{B \times S \times D} \\ \mathbf{Q} &= \mathbf{H}_{\mathbf{G}} \times \mathbf{W}_{\mathbf{q}} \\ \mathbf{K} &= \mathbf{H}_{\mathbf{G}} \times \mathbf{W}_{\mathbf{k}} \end{aligned} \quad (3)$$

where $\mathbf{W}_{\mathbf{q}} \in \mathbb{R}^{D \times D'}$ and $\mathbf{W}_{\mathbf{k}} \in \mathbb{R}^{D \times D'}$ are query and key weight matrices respectively. The subsequent transformer operations remain unchanged, following the standard sequence of cross-attention, feed-forward networks, layer normalization, and residual connections. The value \mathbf{V} matrix maintains its original computation path:

$$\mathbf{V} = \mathbf{H} \times \mathbf{W}_{\mathbf{v}} \quad (4)$$

where $\mathbf{W}_{\mathbf{v}} \in \mathbb{R}^{D \times D'}$ is the value weight matrix. Once \mathbf{Q} , \mathbf{K} and \mathbf{V} are computed, the rest of the transformer encoder (Vaswani et al., 2017) block is unchanged, i.e., cross-attention block followed by feed-forward, layer normalization and residual connection. Figure 2 illustrates our modified BERT architecture with GNN layers (gray shaded area). Multiple GNN layers can be stacked sequentially to enable deeper cross-lingual knowledge transfer.

Strategic Batch Formation for Graph Construction: We propose a batch formation strategy that balances high-resource and low-resource instances while maximizing token overlap between languages. For every batch of size B , we ensure exactly $B/2$ instances from each language domain. Our construction alternates between low-resource and high-resource anchors: we first select a random low-resource instance, then add $(n/2 - 1)$ neighbors from low-resource language and $n/2$ from

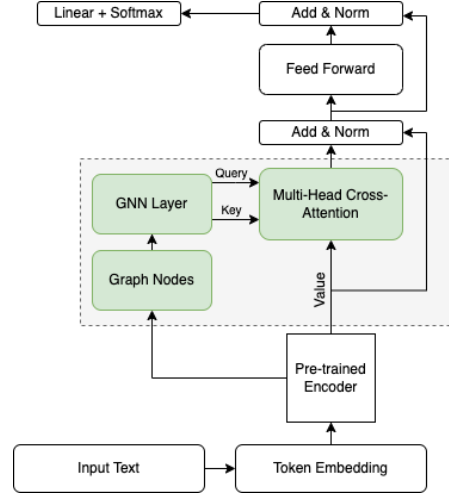


Figure 2: BERT encoder architecture incorporating the GNN layer for cross-lingual knowledge transfer.

high-resource language based on maximum token overlap. These n instances are removed from the available pool to prevent repetition within an epoch. We then select a high-resource anchor and repeat the process, and continue this alternation until the batch is filled. To improve robustness, 70% of the batches follow this strategic formation while the remaining 30% maintain an equal language distribution that selects instances randomly. This prevents over-reliance on specific token patterns while preserving structured knowledge transfer. The process continues across epochs until all low-resource instances are utilized.

During inference, we apply the same principle using training data to form neighborhoods for test instances based on token overlap. This balanced batch construction creates our token interaction graph $G = (T, C)$, enabling effective cross-lingual token relationships without requiring pre-trained resources for the low-resource language.

Cross-Script Edge Construction: When sentences in a batch use different scripts, we establish edges between tokens based on semantic equivalence. We utilize the same translation mechanism described in our TET method (Sec. 3.2) – a combination of dictionary lookup via pymultidictionary (Pizarro, 2025) followed by manual verification.

When two sentences appear in the same batch (e.g., s_1 from English and s_2 from Marathi), and the i -th word of s_1 (w_i^E) conveys similar meaning as the j -th word of s_2 (w_j^M) based on our translation dictionary, connections are established between their tokens. If these words are each represented by single tokens, one edge connects them. If tokenizers split these words into multiple sub-tokens,

then every token of w_i^E will be connected to every token of w_j^M .

For example, if “antarbhasika” (in Marathi) and “cross-lingual” (in English) are present in different sentences of a batch, and they are tokenized as “antar” + “bhasika” and “cross” + “lingual” respectively, edges will be established between all pairs: “antar”-“cross”, “antar”-“lingual”, “bhasika”-“cross” and “bhasika”-“lingual”. While some connections may be noisy, this approach works because: (1) some connections are semantically appropriate (e.g., “antar”-“cross” and “bhasika”-“lingual”), and (2) such cross-token scenarios occur infrequently ($< 20\%$ of connections in a typical batch). During training, the model learns to adjust weights for these connections based on their usefulness for the objective function.

4 Experiments and Results

4.1 Dataset

Our experiments evaluate cross-lingual knowledge transfer across multiple languages and tasks.

For sentiment classification, we employ two high-resource languages: Hindi (Yadav, 2023; Sawant, 2023) and English (Akanksha, 2023), each with 12,000 labeled instances. We use two low-resource target languages: Marathi (Pingle et al., 2023), which shares the Devanagari script with Hindi, and Bangla (Bengali) (Sazzed and Jayarathna, 2019), a language close to Hindi but with its own script. For the Marathi and Bangla datasets, we *simulate* extreme low-resource scenarios with the following random splits: 100 training instances, 1,500 validation instances, and 2,000 test instances. All sentiment classification datasets contain binary labels (positive and negative) with balanced class distributions.

For Named Entity Recognition (NER), we maintain English and Hindi as high-resource languages, with 12,000 training instances (17 entity tags) and 12,084 training instances (13 entity tags) respectively (Jain, 2022; Murthy et al., 2022). For low-resource languages, we use Marathi (Litake et al., 2022) and Malayalam (Mhaske et al., 2023) (which uses a completely different script) with 100 training, 1,500 validation, and 2,000 test instances (14 and 7 unique entity tags for Marathi and Malayalam respectively) created by random sampling.

For the Part-of-Speech (POS) tagging task, we compiled two *truly* low-resource language datasets (Ghosh et al., 2025): Mizo (502 sentences) and

Khasi (507 sentences). Each sentence is fully annotated with POS tags. From both datasets, 100 sentences were randomly selected as the training set and 100 as the validation set; the remainder 302 sentences for Mizo and 307 for Khasi formed the test set. High-resource POS datasets comprise 12,000 labeled sentences each for English (Silveira et al., 2021) and Hindi (Bhat et al., 2025).

4.2 Implementation Details

Following our extreme low-resource assumption, we first used a tinyBERT (Jiao et al., 2020) architecture (without pre-training) to train models from scratch using only 100 labeled instances of each low-resource language, including training new tokenizers. Additionally, we trained two classical ML models such as XGBoost for sentiment classification and CRF for NER and POS tagging tasks. We established several baselines for comparison: (1) *Joint Training (JT)*, which trains on high and low resource languages simultaneously, similar to multilingual models; (2) *JT-HRLAdapLRL*, which sequentially fine-tunes on high-resource data followed by adapter-based fine-tuning on low-resource data; (3) *XLMFT-HRLAdapLRL*, which applies the same sequential approach but starts from XLM-R (Conneau et al., 2020); (4) *LoRA* (Hu et al., 2021), which updates low-rank decomposition matrices instead of full weights; and (5) *AdaMergeX* (Zhao et al., 2025), which combines multiple adaptation methods. For adapter fine-tuning, we used a reduction factor of 16, increasing total parameters by only 1%. For LoRA and AdaMergeX, we followed the recommended parameter settings for encoder tasks from their original papers (Hu et al., 2021; Zhao et al., 2025).

For our high-resource languages, we utilized l3cube-pune/hindi-albert (Joshi, 2022) for Hindi and albert/albert-base-v2 (Lan et al., 2019) for English across both tasks. All experiments were conducted on an Amazon EC2 p4de.24xlarge instance with 8 NVIDIA A100 GPUs (80 GB each), using batch sizes of 128 for most approaches (8 for Scratch Training due to low number of data points, and 120 for GETR methods to accommodate graph construction using 9 neighbors per instance). We employed AdamW with learning rates between $3e-5$ and $3e-7$ for pre-trained models, and $3e-4$ for Scratch Training with TET. Models were trained for 50 epochs with best checkpoints selected via validation loss.

For a fairer comparison between methods, we

carefully balanced parameter counts across all models. Since GETR adds additional GNN layers to the architecture, we removed transformer layers from the pre-trained model to maintain comparable model size. For example, in GETR-GAT, we removed the last 3 transformer encoder layers and added 2 GAT layers, resulting in a parameter count (237,558,024) nearly identical to the Joint Training model (237,557,762). All experiments used the original scripts of the respective languages rather than transliteration. For baseline models, HAL and GETR approaches, we leveraged pre-trained tokenizers from high-resource languages, augmenting them with new tokens from low-resource languages. These newly added tokens were randomly initialized, allowing the model to learn appropriate representations during training

All reported results are evaluated on carefully selected test sets to ensure that there is no overlap with training data (Table 1). As expected, with such limited data and no pre-trained knowledge, Scratch Training models failed to learn meaningful patterns, defaulting to macro-F1 scores of 0.33-0.38 for sentiment classification, and 0.03-0.17 for NER and 0.02-0.16 for POS tagging.

4.3 Results on Sentiment Classification Task

Using English as the high-resource language, the baseline models achieve macro-F1 scores ranging from 0.53 to 0.55 for Marathi, with AdaMergeX performing best among them. Our proposed HAL method with $\alpha = 0.2$ and two layers shows improvement (0.63 with TET), but the GETR-GAT approaches demonstrate substantially greater gains, with GETR-GAT+HAL achieving the best performance (0.75 macro-F1), representing a 20 percentage point improvement over the best baseline. For Bangla as the low-resource language, baseline models achieve macro-F1 scores of 0.63, while GETR-GAT+HAL+TET delivers the best performance at 0.75, a 12 percentage point improvement.

With Hindi as the high-resource language, baseline performance improves significantly (up to 0.76 for AdaMergeX with Marathi), highlighting the benefit of script similarity. When Hindi is used as HRL and Marathi as LRL, TET is not required as they share the Devanagari script, ensuring that Marathi tokens already have pre-trained embeddings from the Hindi model. This explains the absence of TET-based results for this language pair in Table 1. Our HAL approach further boosts performance (0.80 macro-F1), while GETR-

GAT+HAL achieves the highest score for Marathi (0.87 macro-F1), an 11 percentage point improvement over the best baseline. For Hindi-Bangla, GETR-GAT+HAL+TET reaches 0.81 macro-F1, outperforming the best baseline (AdaMergeX at 0.69) by 12 percentage points.

4.4 Results on NER Task

We extended our evaluation to NER for Malayalam and Marathi. Baseline models achieve macro-F1 scores of 0.26–0.28 regardless of high-resource language choice (English or Hindi). GETR-GAT+HAL+TET substantially outperforms baselines: for Malayalam, reaching 0.52 with English (24 pp improvement) and 0.55 with Hindi (27 pp improvement); for Marathi, achieving 0.40 with English (11 pp above best baseline) and 0.44 with Hindi (8 pp above best baseline).

4.5 Results on POS Tagging Task

We evaluate Part-of-Speech tagging on two low-resource languages, Mizo and Khasi. Baseline performance is consistent across both languages and high-resource language choices at 0.71–0.75. GETR-GAT+HAL+TET delivers strong improvements: for Mizo, reaching 0.92 with English and 0.88 with Hindi; for Khasi, achieving 0.88 with English and 0.83 with Hindi, demonstrating the method’s effectiveness on truly low-resource languages with minimal digital presence.

These consistent improvements across different tasks and language families (Indo-Aryan and Dravidian) demonstrate that our GETR approach effectively transfers knowledge regardless of task type or target language. GETR’s superior performance can be attributed to its ability to create dynamic, contextualized connections between tokens across languages, enabling more effective knowledge transfer at a granular level. Unlike static approaches, GETR allows low-resource language tokens to directly incorporate relevant semantic information from high-resource contexts through the graph structure, creating richer representations that better capture cross-lingual patterns. This transfer mechanism operates efficiently through the transformer’s multi-head attention, where Q and K matrices capture the graph-based knowledge of tokens while preserving the original value computations, allowing cross-lingual information to propagate throughout the network. We observed that when using more complex approaches like HAL or GETR, TET’s contribution diminishes. We

| HRL | Method | Sentiment Classification | | NER | | POS Tagging | |
|---------|-----------------------------|--------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| | | Marathi | Bangla | Marathi | Malayalam | Mizo | Khasi |
| - | Scratch Training (tinyBERT) | 0.33 \pm 0.000 | 0.33 \pm 0.000 | 0.03 \pm 0.073 | 0.03 \pm 0.073 | 0.02 \pm 0.061 | 0.04 \pm 0.084 |
| - | Scratch Training (XGBoost) | 0.38 \pm 0.001 | 0.36 \pm 0.001 | - | - | - | - |
| - | Scratch Training (CRF) | - | - | 0.17 \pm 0.004 | 0.15 \pm 0.002 | 0.16 \pm 0.002 | 0.14 \pm 0.003 |
| English | Joint Training | 0.53 \pm 0.002 | 0.63 \pm 0.001 | 0.29 \pm 0.001 | 0.26 \pm 0.002 | 0.75 \pm 0.010 | 0.71 \pm 0.012 |
| | Adapter Finetuning | 0.53 \pm 0.001 | 0.63 \pm 0.002 | 0.29 \pm 0.001 | 0.26 \pm 0.001 | 0.76 \pm 0.008 | 0.72 \pm 0.008 |
| | XLMT-HRLAdapLRL | 0.53 \pm 0.001 | 0.63 \pm 0.002 | 0.29 \pm 0.002 | 0.26 \pm 0.001 | 0.76 \pm 0.009 | 0.72 \pm 0.011 |
| | LoRA | 0.54 \pm 0.001 | 0.63 \pm 0.001 | 0.29 \pm 0.001 | 0.27 \pm 0.001 | 0.78 \pm 0.007 | 0.73 \pm 0.008 |
| | AdaMergeX | 0.55 \pm 0.001 | 0.63 \pm 0.001 | 0.29 \pm 0.001 | 0.28 \pm 0.002 | 0.79 \pm 0.010 | 0.75 \pm 0.007 |
| | HAL | 0.60 \pm 0.001 | 0.64 \pm 0.001 | 0.32 \pm 0.001 | 0.30 \pm 0.003 | 0.85 \pm 0.009 | 0.79 \pm 0.008 |
| | HAL + TET | 0.63 \pm 0.001 | 0.65 \pm 0.003 | 0.33 \pm 0.001 | 0.31 \pm 0.002 | - | - |
| | GETR-GAT | 0.73 \pm 0.001 | 0.72 \pm 0.001 | 0.40 \pm 0.001 | 0.46 \pm 0.001 | 0.91 \pm 0.007 | 0.86 \pm 0.007 |
| | GETR-GAT + TET | 0.74 \pm 0.001 | 0.73 \pm 0.002 | 0.40 \pm 0.001 | 0.47 \pm 0.003 | - | - |
| | GETR-GAT + HAL | 0.75 \pm 0.001 | 0.74 \pm 0.001 | 0.40 \pm 0.001 | 0.51 \pm 0.002 | 0.92 \pm 0.006 | 0.88 \pm 0.006 |
| Hindi | GETR-GAT + HAL + TET | 0.74 \pm 0.001 | 0.75 \pm 0.001 | 0.40 \pm 0.001 | 0.52 \pm 0.001 | - | - |
| | Joint Training | 0.75 \pm 0.004 | 0.67 \pm 0.003 | 0.35 \pm 0.002 | 0.28 \pm 0.002 | 0.75 \pm 0.009 | 0.71 \pm 0.007 |
| | Adapter Finetuning | 0.74 \pm 0.002 | 0.67 \pm 0.002 | 0.34 \pm 0.003 | 0.28 \pm 0.002 | 0.76 \pm 0.009 | 0.71 \pm 0.008 |
| | XLMT-HRLAdapLRL | 0.74 \pm 0.002 | 0.67 \pm 0.002 | 0.34 \pm 0.002 | 0.28 \pm 0.002 | 0.76 \pm 0.008 | 0.72 \pm 0.010 |
| | LoRA | 0.75 \pm 0.001 | 0.68 \pm 0.001 | 0.35 \pm 0.002 | 0.28 \pm 0.001 | 0.77 \pm 0.006 | 0.73 \pm 0.009 |
| | AdaMergeX | 0.76 \pm 0.001 | 0.69 \pm 0.001 | 0.36 \pm 0.001 | 0.28 \pm 0.001 | 0.78 \pm 0.008 | 0.74 \pm 0.007 |
| | HAL | 0.80 \pm 0.005 | 0.72 \pm 0.004 | 0.38 \pm 0.001 | 0.32 \pm 0.003 | 0.83 \pm 0.006 | 0.77 \pm 0.009 |
| | HAL + TET | - | 0.73 \pm 0.002 | - | 0.32 \pm 0.002 | 0.83 \pm 0.005 | 0.77 \pm 0.006 |
| | GETR-GAT | 0.85 \pm 0.001 | 0.79 \pm 0.002 | 0.44 \pm 0.001 | 0.48 \pm 0.001 | 0.88 \pm 0.008 | 0.82 \pm 0.006 |
| | GETR-GAT + TET | - | 0.80 \pm 0.001 | - | 0.49 \pm 0.003 | 0.88 \pm 0.007 | 0.82 \pm 0.006 |
| | GETR-GAT + HAL | 0.87 \pm 0.001 | 0.80 \pm 0.003 | 0.44 \pm 0.001 | 0.53 \pm 0.002 | 0.88 \pm 0.007 | 0.83 \pm 0.007 |
| | GETR-GAT + HAL + TET | - | 0.81 \pm 0.002 | - | 0.55 \pm 0.001 | 0.89 \pm 0.008 | 0.83 \pm 0.009 |

Table 1: Performance (Macro-F1 score) comparison of different training approaches on Sentiment Classification, NER, and POS Tagging tasks when Hindi and English are considered as HRL and Marathi, Bangla, Malayalam, Mizo, and Khasi as LRL. The mean and standard deviation numbers are reported based on 5 independent runs. **Bold** indicates results that are better with statistical significance ($p < 0.005$).

implemented additional baselines including HAL-LRL (augmentation within low-resource language only) and other three XLM-R finetuning variants, all performing comparably to Joint Training. We also evaluated GETR-GCN, but GETR-GAT consistently outperformed it due to GAT’s adaptive edge weighting versus GCN’s equal weighting of connections. Complete results for these experiments appear in Table 4 (in Appendix).

4.6 Cost and Environmental Impact

GETR-GAT incurs modest computational overhead compared to Joint Training while delivering substantial performance gains. Training on AWS p4de.24xlarge instances requires approximately 11% additional time per epoch (≈ 50 minutes vs. ≈ 45 minutes for Joint Training) and 8% more peak GPU memory (≈ 41 GB vs. ≈ 38 GB) due to graph neural network computations and neighborhood construction. Over 50 training epochs, the total training time increases from 37.5 hours to 41.7 hours. Inference is approximately 6.3% slower per sample (0.0101 vs. 0.0095 seconds on p3.2xlarge V100 GPUs). The increased energy consumption (≈ 2.2 kWh vs. ≈ 2.0 kWh per full run) results

in approximately 10% higher CO₂ emissions (≈ 0.99 kg CO₂ vs. ≈ 0.90 kg CO₂), corresponding to roughly 0.09 kg additional CO₂ per model training. These modest computational costs are justified by the significant performance improvements as shown in Table 1. Detailed report is presented in Table 14 (in Appendix).

4.7 Ablation Studies

To evaluate the robustness of our approach and demonstrate its advantage over baseline methods, we compared BhashaSetu (our best-performing GETR-GAT+HAL configuration) with AdaMergeX (Zhao et al., 2025) across varying dataset sizes for NER with Hindi as HRL and Marathi as LRL (Table 2). The results reveal two critical insights. First, with extremely limited low-resource data (10-50 instances), AdaMergeX achieves modest performance (0.05-0.17 F1), while BhashaSetu demonstrates substantially better results even with minimal data, achieving 0.11 F1 with just 10 LRL instances and 0.34 F1 with 50 instances – a 17 percentage points improvement over AdaMergeX at these data scales. The fixed HRL size (12,000) experiment shows BhashaSetu’s

| LRL Size | HRL Size | Macro F1 | |
|----------|----------|------------------|------------------|
| | | AdaMergeX | BhashaSetu |
| 10 | 12000 | 0.05 \pm 0.001 | 0.11 \pm 0.001 |
| 50 | 12000 | 0.17 \pm 0.002 | 0.34 \pm 0.002 |
| 100 | 12000 | 0.35 \pm 0.001 | 0.44 \pm 0.003 |
| 500 | 12000 | 0.39 \pm 0.001 | 0.49 \pm 0.002 |
| 1000 | 12000 | 0.42 \pm 0.002 | 0.52 \pm 0.001 |
| 5000 | 12000 | 0.55 \pm 0.002 | 0.64 \pm 0.003 |
| 10000 | 12000 | 0.71 \pm 0.003 | 0.79 \pm 0.002 |
| 100 | 12000 | 0.35 \pm 0.001 | 0.44 \pm 0.003 |
| 100 | 5000 | 0.22 \pm 0.002 | 0.41 \pm 0.002 |
| 100 | 1000 | 0.11 \pm 0.028 | 0.25 \pm 0.032 |
| 100 | 500 | 0.04 \pm 0.025 | 0.10 \pm 0.023 |

Table 2: NER performance comparison based on Macro-F1 between AdaMergeX and our approach (BhashaSetu) with Hindi as high-resource and Marathi as low-resource language under varying dataset sizes.

consistent advantage across all LRL sizes, with improvements of 9-17 percentage points, though the relative gap narrows as low-resource data increases.

The second experiment, keeping LRL fixed at 100 instances while varying HRL size, reveals that AdaMergeX’s performance degrades dramatically with decreasing HRL data (from 0.35 F1 with 12,000 instances to just 0.04 F1 with 500 instances). While BhashaSetu also shows decreased performance with less HRL data, it maintains substantially better results (0.10 F1 even with just 500 HRL instances) and demonstrates greater resilience to HRL data reduction. These results highlight both BhashaSetu’s effectiveness at enabling cross-lingual knowledge transfer and its superior ability to leverage limited high-resource data compared to AdaMergeX. Our additional experiments on sentiment classification (details in Tables 7 and 8 in Appendix) reinforce these findings, with BhashaSetu outperforming AdaMergeX by 14-28 percentage points for Hindi-Bangla and 12-27 percentage points for English-Bangla pairs across various dataset sizes.

We investigated whether increasing batch size during training improves training efficiency while maintaining performance. Training GETR-GAT with batch sizes of 60, 120, and 4096 on sentiment classification tasks across multiple language pairs, we found that smaller batches (120) achieved comparable performance to large batch sizes (4096) but required approximately 14% more training steps to saturation. This is expected, as larger batches enable more direct token connections between low-resource and high-resource languages through the graph structure, allowing the model to learn global semantics more efficiently. Smaller batches re-

quire successive approximations through the 30% random sampling strategy to establish these cross-lingual connections across epochs. These results demonstrate that GETR’s effectiveness is robust to different batch sizes, though computational constraints typically limit practical batch sizes. Detailed results are presented in Sec. C.1 of Appendix.

To assess GETR’s resilience to incomplete or noisy cross-lingual edge information, we conducted experiments by systematically reducing the percentage of cross-lingual token connections in the graph. Each typical batch contains 600–1000 token connections established through our translation mappings. We evaluated GETR-GAT performance with 100%, 70%, 50%, 30%, and 0% of these connections, where 0% reduces to standard Joint Training. Performance degraded gracefully with fewer connections, demonstrating that GETR maintains substantial gains even with 70% of the edges (typically within 2–5 percentage points of full performance) and converges toward Joint Training baseline at 0% edge retention. This robustness suggests that GETR does not require perfect bilingual lexicons and can function effectively even with partial or noisy translation dictionaries. Detailed results are presented in Sec. C.2 of Appendix.

5 Conclusions

In this paper, we addressed the challenge of cross-lingual knowledge transfer for extreme low-resource scenarios. We proposed two adopted baselines, HAL and TET, and a novel GETR mechanism. Experimental results demonstrate that while traditional multilingual and cross-lingual models struggle with extreme data scarcity, our proposed approaches effectively leverage knowledge from high-resource languages.

Future work includes exploring linguistic insights and self-supervised pre-training strategies specific to low-resource languages, memory-optimized implementations of graph neural networks, and cross-lingual transfer for a wider range of tasks and language pairs.

Acknowledgements

Following ARR’s AI Writing/Coding Assistance Policy, we acknowledge using Claude 3.7 Sonnet (Anthropic, 2025) for editorial assistance while maintaining full ownership of and responsibilities for all scientific contributions.

Limitations

While our proposed approaches demonstrate strong performance across different tasks and language pairs, we acknowledge certain aspects that present opportunities for future research. Our experiments primarily focus on Indian languages from both Indo-Aryan and Dravidian families, which could be extended to typologically more distant language pairs with different word orders or morphological systems in future work.

Although BhashaSetu is effective with minimal low-resource data (100 instances), we observe that transfer performance correlates with high-resource language data availability, a common pattern in transfer learning approaches. This relationship between source data volume and transfer effectiveness presents an interesting direction for developing more data-efficient transfer techniques.

The Token Embedding Transfer approach benefits from word-level translation capabilities between language pairs. While such resources exist for many language combinations, future work could explore unsupervised methods for establishing cross-lingual correspondences when traditional bilingual dictionaries are unavailable.

Our Graph-Enhanced Token Representation approach introduces additional computational complexity during training and inference due to graph construction operations and GNN computations compared to simpler methods. However, this computational investment delivers substantially improved performance (21-27 percentage points gain in F1 scores), representing a favorable trade-off in many practical scenarios. Future implementations could explore optimization techniques to reduce this overhead.

Finally, while we demonstrate effectiveness on classification tasks (sentiment analysis and NER), extending these approaches to generative tasks involving neural machine translation or summary generation represents a promising direction for future research. This would further validate the versatility of our framework across the broader NLP task spectrum.

Ethics Statement

This research aims to promote linguistic inclusivity by addressing the technological disparity between high-resource and low-resource languages. We acknowledge that NLP capabilities have predominantly benefited widely-spoken languages, po-

tentially exacerbating digital divides along linguistic lines. All datasets used in our experiments are publicly available with appropriate citations, and we did not collect or annotate new data that might introduce privacy concerns.

We recognize that transfer learning approaches may inadvertently propagate biases from source to target languages; however, our work takes a step toward mitigating representation disparities by enabling better performance with minimal labeled data in low-resource languages. Due to the focus on extremely low-resource settings (approximately 100 training instances), the computational requirements for target language adaptation were substantially lower than those typically needed for high-resource language model development, reducing the environmental impact compared to training large language models from scratch. While the GETR approaches do introduce additional computational overhead during the knowledge transfer process, the overall resource consumption remains modest relative to pre-training large multilingual models. This efficiency is particularly beneficial for researchers and practitioners with limited computational resources working on low-resource language technologies.

While we focused on Indian languages in this study, we believe that similar approaches could benefit other low-resource languages globally, contributing to more equitable language technology development. We emphasize that the performance improvements demonstrated should be considered within the context of the limitations described in our paper, and that practical applications would require careful consideration of cultural and linguistic nuances specific to each target community.

References

- Akanksha. 2023. Sentiment analysis dataset. <https://www.kaggle.com/code/akanksha10/sentiment-analysis-dataset>. Accessed: 2024.
- Anthropic. 2025. Claude 3.7 sonnet [large language model]. <https://www.anthropic.com>. Released February 24, 2025. Accessed 2025-05-19.
- R. A. Bhat, R. Bhatt, A. Farudi, P. Klassen, B. Narasimhan, M. Palmer, O. Rambow, D. M. Sharma, A. Vaidya, S. R. Vishnu, and F. Xia. 2025. The Hindi/Urdu treebank project. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*. Springer Press. To be updated with actual year.

- Federico Cassano, John Gouwar, Francesca Lucchetti, Claire Schlesinger, Anders Freeman, Carolyn Jane Anderson, Molly Q Feldman, Michael Greenberg, Abhinav Jangda, and Arjun Guha. 2024. [Knowledge transfer from high-resource to low-resource programming languages for code llms](#). *Proc. ACM Program. Lang.*, 8(OOPSLA2).
- Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. [Mix-Text: Linguistically-informed interpolation of hidden space for semi-supervised text classification](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Jiequan Cui, Zhuotao Tian, Zhisheng Zhong, Xiaojuan Qi, Bei Yu, and Hanwang Zhang. 2024. Decoupled kullback-leibler divergence loss. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*, Red Hook, NY, USA. Curran Associates Inc.
- C. M. Downey, Terra Blevins, Dhvani Serai, Dwija Parikh, and Shane Steinert-Threlkeld. 2024. [Targeted multilingual adaptation for low-resource language families](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15647–15663, Miami, Florida, USA. Association for Computational Linguistics.
- Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Edouard Hovy. 2021. [A survey of data augmentation approaches for NLP](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.
- Soumyadip Ghosh, Henry Lalsiam, Dorothy Marbaniang, Gracious Temsen, Rahul Mishra, and Parameswari Krishnamurthy. 2025. [Towards resource-rich Mizo and Khasi in NLP: Resource development, synthetic data generation and model building](#). In *Proceedings of [Conference Name]*, pages 228–239.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. [Parameter-Efficient Transfer Learning for NLP](#). *arXiv e-prints*, arXiv:1902.00751.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [LoRA: Low-Rank Adaptation of Large Language Models](#). *arXiv e-prints*, arXiv:2106.09685.
- Naman Jain. 2022. Ner dataset. <https://www.kaggle.com/datasets/namanj27/ner-dataset>. Retrieved May 17, 2025.
- Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. [TinyBERT: Distilling BERT for natural language understanding](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.
- Raviraj Joshi. 2022. [L3Cube-HindBERT and DevBERT: Pre-Trained BERT Transformer models for Devanagari based Hindi and Marathi Languages](#). *arXiv e-prints*, arXiv:2211.11418.
- Thomas N. Kipf and Max Welling. 2016. [Semi-Supervised Classification with Graph Convolutional Networks](#). *arXiv e-prints*, arXiv:1609.02907.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. [ALBERT: A Lite BERT for Self-supervised Learning of Language Representations](#). *arXiv e-prints*, arXiv:1909.11942.
- Seungyeon Lee, Seongtae Hong, Hyeonseok Moon, and Heuiseok Lim. 2025. [Semantic aware linear transfer by recycling pre-trained language models for cross-lingual transfer](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16180–16193, Vienna, Austria. Association for Computational Linguistics.
- Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabza. 2023. [XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics.
- Onkar Litake, Maithili Ravindra Sabane, Parth Sachin Patil, Aparna Abhijeet Ranade, and Raviraj Joshi. 2022. [L3Cube-MahaNER: A Marathi named entity recognition dataset and BERT models](#). In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 29–34, Marseille, France. European Language Resources Association.
- Yihong Liu, Haotian Ye, Leonie Weissweiler, Renhao Pei, and Hinrich Schuetze. 2023. [Crosslingual transfer learning for low-resource languages based on multilingual colexification graphs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8376–8401, Singapore. Association for Computational Linguistics.

- Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2023. Cross-entropy loss functions: theoretical analysis and applications. In *Proceedings of the 40th International Conference on Machine Learning, ICML'23*. JMLR.org.
- Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. *mmBERT: A Modern Multilingual Encoder with Annealed Language Learning*. *arXiv e-prints*, arXiv:2509.06888.
- Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. 2023. *Naamapadam: A large-scale named entity annotated data for Indic languages*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10441–10456, Toronto, Canada. Association for Computational Linguistics.
- Rudra Murthy, Pallab Bhattacharjee, Rahul Sharnagat, Jyotsana Khatri, Diptesh Kanojia, and Pushpak Bhat-tacharyya. 2022. *HiNER: A large Hindi named entity recognition dataset*. In *Proceedings of the Thirtieth Language Resources and Evaluation Conference*, pages 4467–4476, Marseille, France. European Language Resources Association.
- Long H. B. Nguyen, Binh Nguyen, Binh Le, and Dien Dinh. 2023. *Exploring graph-based transformer encoder for low-resource neural machine translation*. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* Just Accepted.
- Aabha Pingle, Aditya Vyawahare, Isha Joshi, Rahul Tangsali, and Raviraj Joshi. 2023. *L3Cube-MahaSent-MD: A multi-domain Marathi sentiment analysis dataset and transformer models*. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 274–281, Hong Kong, China. Association for Computational Linguistics.
- Pablo Pizarro. 2025. Pymultidictionary: Multilingual dictionary module for python. <https://github.com/ppizarro/PyMultiDictionary>. Version 5.1.0, Accessed: 2025-07-28.
- François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. 2024. *Trans-Tokenization and Cross-lingual Vocabulary Transfers: Language Adaptation of LLMs for Low-Resource NLP*. *arXiv e-prints*, arXiv:2408.04303.
- Onkar Sawant. 2023. Hindi sentiment analysis dataset. <https://www.kaggle.com/datasets/onkarsawant5613/hindi-sentiment-analysis>. Accessed: 2024.
- Salim Sazed and Sampath Jayarathna. 2019. *A sentiment classification in Bengali and machine translated english corpus*. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 107–114.
- Naiara Silveira, Timothy Dozat, Stephan Schuster, Matthew Connor, Marie-Catherine de Marneffe, Nathan Schneider, Eric Chi, Samuel Bowman, Christopher Manning, Hao Zhu, Danielle Galbraith, and John Bauer. 2021. Universal dependencies english web treebank (ud-ewt). Linguistic Data Consortium. [Dataset].
- Lichao Sun, Congying Xia, Wenpeng Yin, Tingting Liang, Philip Yu, and Lifang He. 2020. *Mixup-transformer: Dynamic data augmentation for NLP tasks*. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3436–3440, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. *Attention is all you need*. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. *Graph Attention Networks*. *arXiv e-prints*, arXiv:1710.10903.
- Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, Aaron Courville, David Lopez-Paz, and Yoshua Bengio. 2018. *Manifold Mixup: Better Representations by Interpolating Hidden States*. *arXiv e-prints*, arXiv:1806.05236.
- Ziyun Wang, Xuan Liu, Peiji Yang, Shixing Liu, and Zhisheng Wang. 2021. *Cross-lingual text classification with heterogeneous graph neural network*. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 612–620, Online. Association for Computational Linguistics.
- Shijie Wu and Mark Dredze. 2020. *Are all languages created equal in multilingual BERT?* In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.
- Siddharth Yadav. 2023. *Hindi sentiment analysis*. https://github.com/sid573/Hindi_Sentiment_Analysis.
- Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2017. *mixup: Beyond Empirical Risk Minimization*. *arXiv e-prints*, arXiv:1710.09412.
- Xinghua Zhang, Bowen Yu, Yubin Wang, Tingwen Liu, Taoyu Su, and Hongbo Xu. 2022. *Exploring modular task decomposition in cross-domain named entity recognition*. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22*, page 301–311, New York, NY, USA. Association for Computing Machinery.

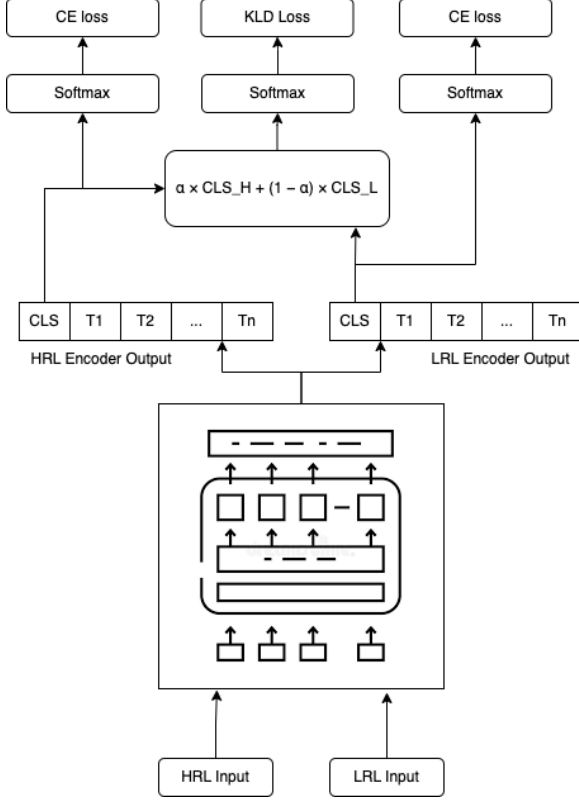


Figure 3: Architecture incorporating the Hidden Augmentation Layer (HRL and LRL inputs are high- and low-resource language inputs respectively)

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2025. [AdaMergeX: Cross-lingual transfer with large language models via adaptive adapter merging](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9785–9800, Albuquerque, New Mexico. Association for Computational Linguistics.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. [Graph neural networks: A review of methods and applications](#). *AI Open*, 1:57–81.

Appendix

A HAL

Figure 3 illustrates our modified architecture incorporating the hidden augmentation layer. The framework can be further extended by adding multiple transformer layers above E_M and performing augmentation at each layer’s CLS output, thus enabling hierarchical knowledge fusion.

B Dynamic Mixing Coefficient Analysis

We experimented with a dynamic mixing coefficient that changes linearly during model training, transitioning from 1 (favoring the high-resource language) to 0 (favoring the low-resource language). The definition of dynamic α is given by:

$$\alpha = \frac{1 - s}{T}$$

where s represents the number of training steps and T is the total number of training steps. This approach aims to gradually shift the model’s focus from leveraging high-resource language knowledge during early training phases to increasingly relying on low-resource language-specific patterns as training progresses.

Implementing this dynamic mixing strategy, we found no major improvement over the existing HAL approach. Table 3 presents the comprehensive results for all high-resource and low-resource language pairs across both Sentiment Classification and Named Entity Recognition tasks, comparing the original HAL with fixed $\alpha = 0.2$ to HAL with the dynamic α (HAL-DynamicAlpha).

B.1 Discussion

The empirical results demonstrate that the dynamic mixing coefficient strategy yields negligible performance differences compared to the fixed $\alpha = 0.2$ configuration. Across all language pairs and both tasks, the performance metrics remain nearly identical between HAL and HAL-DynamicAlpha, with overlapping confidence intervals. This finding indicates that the computational overhead introduced by the dynamic schedule does not translate into meaningful improvements in model performance.

The stability of our results across the dynamic schedule further validates the robustness of the HAL approach with a fixed mixing coefficient of 0.2, as reported in the main paper. The dynamic strategy was motivated by the hypothesis that gradually shifting focus from high-resource to low-resource language knowledge during training might prevent negative transfer and improve generalization. However, our experiments suggest that a carefully tuned fixed mixing coefficient is sufficient to balance knowledge transfer from the high-resource language with preservation of low-resource language characteristics, making the added complexity of dynamic scheduling unnecessary. Note that while we implemented and evalu-

| HRL | Method | Sentiment Classification | | NER | |
|---------|------------------|--------------------------|------------------|------------------|------------------|
| | | Marathi as LRL | Bangla as LRL | Marathi as LRL | Malayalam as LRL |
| English | HAL | 0.60 ± 0.001 | 0.64 ± 0.001 | 0.32 ± 0.001 | 0.30 ± 0.003 |
| | HAL-DynamicAlpha | 0.60 ± 0.002 | 0.64 ± 0.002 | 0.32 ± 0.001 | 0.30 ± 0.002 |
| Hindi | HAL | 0.80 ± 0.005 | 0.72 ± 0.004 | 0.38 ± 0.001 | 0.32 ± 0.003 |
| | HAL-DynamicAlpha | 0.80 ± 0.004 | 0.72 ± 0.003 | 0.38 ± 0.001 | 0.32 ± 0.003 |

Table 3: Performance (Macro-F1 score) comparison of HAL and HAL-DynamicAlpha on Sentiment Classification and NER tasks across different high-resource and low-resource language pairs. Results are reported with mean and standard deviation based on independent runs.

ated a linear schedule for the dynamic mixing coefficient, more adaptive scheduling strategies such as uncertainty-based, performance-driven, or non-linear curricula were not explored in this work due to time and scope constraints. This clarifies the boundaries of our investigation for readers and future follow-up studies.

B.2 Results on Sentiment Classification Task and NER

We extensively evaluated our approach against multiple baselines, including parameter-efficient fine-tuning methods and XLM-R variants. For XLM-R, we tested: (1) XLMFT-LRL: fine-tuning only on low-resource data; (2) XLMFT-HRLRL: joint fine-tuning on both language datasets; (3) XLMFT-HRL2LRL: sequential fine-tuning on high-resource followed by low-resource data; and (4) XLMFT-HRLAdapLRL: fine-tuning on high-resource data followed by adapter-based fine-tuning on low-resource data with frozen base weights. Additionally, we evaluated LoRA and AdaMergeX as parameter-efficient alternatives, and HAL-LRL which applies augmentation only within the low-resource language. Our results show XLM-R variants perform comparably to Joint Training across all configurations, while HAL-LRL shows no improvement over Joint Training due to limited augmentation diversity in the extremely small low-resource dataset.

To understand the impact of mixing coefficient α in Hidden Augmentation Layer (HAL), we conducted experiments with different α values ranging from 0.1 to 0.8 (Table 5). For both English and Hindi as high-resource languages, $\alpha=0.2$ yields the best performance, achieving accuracy/F1 scores of 0.610/0.590 and 0.860/0.860 respectively. The performance gradually degrades as α increases, with a more pronounced decline after $\alpha=0.5$. This suggests that while knowledge from the high-resource

language provides useful linguistic patterns and Sentiment structures, excessive reliance on it diminishes the model’s ability to capture the unique characteristics and nuances of the low-resource language. The optimal performance at $\alpha=0.2$ indicates that a balanced approach, where the model primarily learns from the low-resource language while leveraging complementary features from the high-resource language, is most effective. Notably, even with declining performance at higher α values, the model maintains reasonable performance (minimum accuracy of 0.590 for English and 0.830 for Hindi as HRL), indicating the robustness of the HAL approach across different mixing ratios.

We analyzed the impact of HAL depth by varying the number of layers from 1 to 6 (Table 6). For both English and Hindi as high-resource languages, 2 HAL layers yield optimal performance (accuracy/F1: 0.610/0.590 and 0.860/0.860 respectively), with secondary peaks at depth 4 for English (0.598/0.582) and depth 5 for Hindi (0.848/0.845), suggesting that while multiple HAL layers aid in knowledge transfer, excessive depth might lead to over-abstraction of features. Similarly, for both GETR-GCN and GETR-GAT approaches, three GNN layers demonstrated the best performance on the test set metrics, indicating an optimal depth for graph-based token interaction.

We extended our robustness evaluation to sentiment classification with Bangla as the low-resource language, testing both Hindi and English as high-resource languages (Table 7). The results reveal consistent advantages for BhashaSetu across all data configurations. With minimal low-resource data (10 instances), Joint Training achieves only 0.33 macro-F1 for both HRLs, while BhashaSetu reaches 0.61 with Hindi and 0.60 with English—an approximately 85% improvement. This advantage persists across all LRL sizes, though the gap narrows as training data increases. Hindi consistently

| HRL | Method | Sentiment Classification | | NER | |
|---------|----------------------|------------------------------------|------------------------------------|------------------------------------|------------------------------------|
| | | Marathi as LRL | Bangla as LRL | Marathi as LRL | Malayalam as LRL |
| - | Scratch Training | 0.33 \pm 0.000 | 0.33 \pm 0.000 | 0.03 \pm 0.073 | 0.03 \pm 0.073 |
| English | Joint Training | 0.53 \pm 0.002 | 0.63 \pm 0.001 | 0.29 \pm 0.001 | 0.26 \pm 0.002 |
| | Adapter Finetuning | 0.53 \pm 0.001 | 0.63 \pm 0.002 | 0.29 \pm 0.001 | 0.26 \pm 0.001 |
| | XMLFT-LRL | 0.49 \pm 0.002 | 0.60 \pm 0.004 | 0.27 \pm 0.005 | 0.23 \pm 0.003 |
| | XMLFT-HRLRL | 0.53 \pm 0.002 | 0.63 \pm 0.001 | 0.29 \pm 0.001 | 0.26 \pm 0.001 |
| | XMLFT-HRL2LRL | 0.52 \pm 0.002 | 0.63 \pm 0.004 | 0.28 \pm 0.003 | 0.24 \pm 0.003 |
| | XMLFT-HRLAdapLRL | 0.53 \pm 0.001 | 0.63 \pm 0.002 | 0.29 \pm 0.002 | 0.26 \pm 0.001 |
| | LoRA | 0.54 \pm 0.001 | 0.63 \pm 0.001 | 0.29 \pm 0.001 | 0.27 \pm 0.001 |
| | AdaMergeX | 0.55 \pm 0.001 | 0.63 \pm 0.001 | 0.29 \pm 0.001 | 0.28 \pm 0.002 |
| | HAL-LRL | 0.52 \pm 0.002 | 0.63 \pm 0.001 | 0.29 \pm 0.001 | 0.26 \pm 0.001 |
| | HAL | 0.60 \pm 0.001 | 0.64 \pm 0.001 | 0.32 \pm 0.001 | 0.30 \pm 0.003 |
| | HAL + TET | 0.63 \pm 0.001 | 0.65 \pm 0.003 | 0.33 \pm 0.001 | 0.31 \pm 0.002 |
| | GETR-GCN | 0.69 \pm 0.002 | 0.68 \pm 0.001 | 0.36 \pm 0.001 | 0.37 \pm 0.001 |
| | GETR-GCN + TET | 0.68 \pm 0.001 | 0.69 \pm 0.003 | 0.36 \pm 0.001 | 0.37 \pm 0.002 |
| | GETR-GCN + HAL | 0.70 \pm 0.001 | 0.70 \pm 0.002 | 0.39 \pm 0.001 | 0.43 \pm 0.003 |
| | GETR-GCN + HAL + TET | 0.70 \pm 0.002 | 0.70 \pm 0.001 | 0.39 \pm 0.001 | 0.43 \pm 0.002 |
| | GETR-GAT | 0.73 \pm 0.001 | 0.72 \pm 0.001 | 0.40 \pm 0.001 | 0.46 \pm 0.001 |
| | GETR-GAT + TET | 0.74 \pm 0.001 | 0.73 \pm 0.002 | 0.40 \pm 0.001 | 0.47 \pm 0.003 |
| | GETR-GAT + HAL | 0.75 \pm 0.001 | 0.74 \pm 0.001 | 0.40 \pm 0.001 | 0.51 \pm 0.002 |
| | GETR-GAT + HAL + TET | 0.74 \pm 0.001 | 0.75 \pm 0.001 | 0.40 \pm 0.001 | 0.52 \pm 0.001 |
| Hindi | Joint Training | 0.75 \pm 0.004 | 0.67 \pm 0.003 | 0.35 \pm 0.002 | 0.28 \pm 0.002 |
| | Adapter Finetuning | 0.74 \pm 0.002 | 0.67 \pm 0.002 | 0.34 \pm 0.003 | 0.28 \pm 0.002 |
| | XMLFT-LRL | 0.71 \pm 0.003 | 0.62 \pm 0.005 | 0.30 \pm 0.004 | 0.26 \pm 0.004 |
| | XMLFT-HRLRL | 0.75 \pm 0.001 | 0.67 \pm 0.001 | 0.34 \pm 0.001 | 0.28 \pm 0.001 |
| | XMLFT-HRL2LRL | 0.75 \pm 0.003 | 0.67 \pm 0.004 | 0.34 \pm 0.004 | 0.27 \pm 0.001 |
| | XMLFT-HRLAdapLRL | 0.74 \pm 0.002 | 0.67 \pm 0.002 | 0.34 \pm 0.003 | 0.28 \pm 0.002 |
| | LoRA | 0.75 \pm 0.001 | 0.68 \pm 0.001 | 0.35 \pm 0.002 | 0.28 \pm 0.001 |
| | AdaMergeX | 0.76 \pm 0.001 | 0.69 \pm 0.001 | 0.36 \pm 0.001 | 0.28 \pm 0.001 |
| | HAL-LRL | 0.75 \pm 0.003 | 0.67 \pm 0.002 | 0.35 \pm 0.001 | 0.27 \pm 0.001 |
| | HAL | 0.80 \pm 0.005 | 0.72 \pm 0.004 | 0.38 \pm 0.001 | 0.32 \pm 0.003 |
| | HAL + TET | - | 0.73 \pm 0.002 | - | 0.32 \pm 0.002 |
| | GETR-GCN | 0.82 \pm 0.001 | 0.75 \pm 0.001 | 0.42 \pm 0.002 | 0.38 \pm 0.001 |
| | GETR-GCN + TET | - | 0.75 \pm 0.002 | - | 0.38 \pm 0.002 |
| | GETR-GCN + HAL | 0.83 \pm 0.002 | 0.76 \pm 0.001 | 0.43 \pm 0.001 | 0.44 \pm 0.003 |
| | GETR-GCN + HAL + TET | - | 0.76 \pm 0.002 | - | 0.44 \pm 0.002 |
| | GETR-GAT | 0.85 \pm 0.001 | 0.79 \pm 0.002 | 0.44 \pm 0.001 | 0.48 \pm 0.001 |
| | GETR-GAT + TET | - | 0.80 \pm 0.001 | - | 0.49 \pm 0.003 |
| | GETR-GAT + HAL | 0.87 \pm 0.001 | 0.80 \pm 0.003 | 0.44 \pm 0.001 | 0.53 \pm 0.002 |
| | GETR-GAT + HAL + TET | - | 0.81 \pm 0.002 | - | 0.55 \pm 0.001 |

Table 4: Performance (Macro-F1 score) comparison of different training approaches on sentiment classification and NER datasets when Hindi and English are considered as HRL and Marathi, Bangla and Malayalam as LRL. The mean and standard deviation numbers are reported based on 5 independent runs.

outperforms English as the high-resource language, with BhashaSetu reaching 0.94 F1 using Hindi versus 0.89 F1 using English at 8,000 LRL instances.

The fixed LRL experiments (100 instances) with varying HRL size reveal BhashaSetu’s remarkable resilience to limited high-resource data. With just 500 HRL instances, BhashaSetu maintains 0.62 F1 (Hindi) and 0.57 F1 (English), while Joint Training drops to 0.43 and 0.41 respectively. Most impressively, BhashaSetu with just 1,000 Hindi instances (0.73 F1) outperforms Joint Training with the full 12,000 instances (0.67 F1). These results demonstrate BhashaSetu’s exceptional data efficiency in leveraging limited resources for cross-lingual transfer and confirm its effectiveness across both NER

and sentiment classification tasks, regardless of the specific high-resource language used.

To evaluate the robustness of our approach on sentiment classification, we conducted extensive experiments varying dataset sizes with both Hindi and English as high-resource languages for Bangla (Table 7). With Hindi as HRL, BhashaSetu demonstrates remarkable effectiveness, achieving 0.61 macro-F1 with just 10 LRL instances compared to Joint Training’s 0.33—an improvement of 28 percentage points. This advantage persists as LRL size increases, maintaining improvements of 12-21 percentage points up to 8,000 instances (the maximum available in our Bangla dataset), where BhashaSetu achieves 0.94 macro-F1 compared to

| HRL | LRL | α | Metrics | |
|---------|---------|------------|-----------------------------------|-----------------------------------|
| | | | Accuracy | F1 |
| English | Marathi | 0.1 | 0.602 \pm 0.004 | 0.582 \pm 0.005 |
| | | 0.2 | 0.610\pm0.004 | 0.590\pm0.005 |
| | | 0.3 | 0.605 \pm 0.003 | 0.578 \pm 0.004 |
| | | 0.4 | 0.598 \pm 0.004 | 0.571 \pm 0.005 |
| | | 0.5 | 0.595 \pm 0.005 | 0.565 \pm 0.004 |
| | | 0.6 | 0.592 \pm 0.004 | 0.558 \pm 0.005 |
| | | 0.7 | 0.591 \pm 0.005 | 0.552 \pm 0.004 |
| | | 0.8 | 0.590 \pm 0.004 | 0.550 \pm 0.005 |
| Hindi | Marathi | 0.1 | 0.852 \pm 0.004 | 0.848 \pm 0.005 |
| | | 0.2 | 0.860\pm0.003 | 0.860\pm0.005 |
| | | 0.3 | 0.848 \pm 0.004 | 0.845 \pm 0.004 |
| | | 0.4 | 0.842 \pm 0.005 | 0.840 \pm 0.005 |
| | | 0.5 | 0.838 \pm 0.004 | 0.835 \pm 0.004 |
| | | 0.6 | 0.834 \pm 0.005 | 0.832 \pm 0.005 |
| | | 0.7 | 0.832 \pm 0.004 | 0.831 \pm 0.004 |
| | | 0.8 | 0.830 \pm 0.005 | 0.830 \pm 0.005 |

Table 5: Performance comparison of HAL approach with different high-resource languages and varying α values. HRL: High Resource Language, LRL: Low Resource Language

Joint Training’s 0.82.

Similar patterns emerge with English as HRL, though with slightly lower absolute performance due to script differences. BhashaSetu achieves 0.60 macro-F1 with 10 LRL instances (27 percentage points over Joint Training) and maintains substantial improvements through 8,000 instances (0.89 vs 0.78 macro-F1). The fixed LRL experiments (100 instances) reveal BhashaSetu’s superior resilience to HRL data reduction: with Hindi, performance drops from 0.81 to 0.62 macro-F1 as HRL size decreases from 12,000 to 500, while Joint Training falls more sharply from 0.67 to 0.43. English shows similar trends, with BhashaSetu maintaining better performance (0.75 to 0.57) compared to Joint Training’s steeper decline (0.63 to 0.41).

| HRL | LRL | HAL Depth | Metrics | |
|---------|---------|-----------|-----------------------------------|-----------------------------------|
| | | | Accuracy | F1 |
| English | Marathi | 1 | 0.592 \pm 0.004 | 0.575 \pm 0.005 |
| | | 2 | 0.610\pm0.004 | 0.590\pm0.005 |
| | | 3 | 0.588 \pm 0.003 | 0.562 \pm 0.004 |
| | | 4 | 0.598 \pm 0.004 | 0.582 \pm 0.005 |
| | | 5 | 0.575 \pm 0.005 | 0.545 \pm 0.004 |
| | | 6 | 0.570 \pm 0.004 | 0.540 \pm 0.005 |
| Hindi | Marathi | 1 | 0.842 \pm 0.004 | 0.838 \pm 0.005 |
| | | 2 | 0.860\pm0.003 | 0.860\pm0.005 |
| | | 3 | 0.835 \pm 0.004 | 0.832 \pm 0.004 |
| | | 4 | 0.825 \pm 0.005 | 0.818 \pm 0.005 |
| | | 5 | 0.848 \pm 0.004 | 0.845 \pm 0.004 |
| | | 6 | 0.810 \pm 0.005 | 0.800 \pm 0.005 |

Table 6: Impact of HAL depth on model performance. HRL: High Resource Language, LRL: Low Resource Language

| HRL | HRL Size | LRL Size | Macro F1 + JT | Macro F1 + BhashaSetu |
|---|----------|----------|------------------|-----------------------|
| <i>Fixed HRL Size, Varying LRL Size</i> | | | | |
| Hindi | 12000 | 10 | 0.33 \pm 0.001 | 0.61 \pm 0.001 |
| Hindi | 12000 | 50 | 0.51 \pm 0.002 | 0.72 \pm 0.002 |
| Hindi | 12000 | 100 | 0.67 \pm 0.001 | 0.81 \pm 0.003 |
| Hindi | 12000 | 500 | 0.69 \pm 0.001 | 0.83 \pm 0.002 |
| Hindi | 12000 | 1000 | 0.73 \pm 0.002 | 0.87 \pm 0.001 |
| Hindi | 12000 | 5000 | 0.79 \pm 0.002 | 0.92 \pm 0.003 |
| Hindi | 12000 | 8000 | 0.82 \pm 0.003 | 0.94 \pm 0.002 |
| English | 12000 | 10 | 0.33 \pm 0.001 | 0.60 \pm 0.001 |
| English | 12000 | 50 | 0.49 \pm 0.002 | 0.68 \pm 0.002 |
| English | 12000 | 100 | 0.63 \pm 0.001 | 0.75 \pm 0.003 |
| English | 12000 | 500 | 0.65 \pm 0.001 | 0.78 \pm 0.002 |
| English | 12000 | 1000 | 0.69 \pm 0.002 | 0.81 \pm 0.001 |
| English | 12000 | 5000 | 0.74 \pm 0.002 | 0.87 \pm 0.003 |
| English | 12000 | 8000 | 0.78 \pm 0.003 | 0.89 \pm 0.002 |
| <i>Fixed LRL Size, Varying HRL Size</i> | | | | |
| Hindi | 12000 | 100 | 0.67 \pm 0.001 | 0.81 \pm 0.003 |
| Hindi | 5000 | 100 | 0.61 \pm 0.002 | 0.76 \pm 0.002 |
| Hindi | 1000 | 100 | 0.52 \pm 0.023 | 0.73 \pm 0.003 |
| Hindi | 500 | 100 | 0.43 \pm 0.022 | 0.62 \pm 0.006 |
| English | 12000 | 100 | 0.63 \pm 0.001 | 0.75 \pm 0.003 |
| English | 5000 | 100 | 0.55 \pm 0.002 | 0.71 \pm 0.002 |
| English | 1000 | 100 | 0.50 \pm 0.023 | 0.65 \pm 0.003 |
| English | 500 | 100 | 0.41 \pm 0.022 | 0.57 \pm 0.006 |

Table 7: Sentiment Classification performance comparison based on Macro-F1 between Joint Training (JT) and our approach (BhashaSetu) with Hindi and English as high-resource and Bangla as low-resource language under varying dataset sizes.

These results demonstrate BhashaSetu’s effectiveness across different data regimes and language pairs, with particularly strong performance when languages share scripts.

| LRL Size | HRL Size | Macro F1 + AdaMergeX | Macro F1 + BhashaSetu |
|---|----------|----------------------|-----------------------|
| <i>Fixed HRL Size, Varying LRL Size</i> | | | |
| 10 | 12000 | 0.02 \pm 0.001 | 0.11 \pm 0.001 |
| 50 | 12000 | 0.13 \pm 0.002 | 0.34 \pm 0.002 |
| 100 | 12000 | 0.29 \pm 0.001 | 0.40 \pm 0.003 |
| 500 | 12000 | 0.34 \pm 0.001 | 0.46 \pm 0.002 |
| 1000 | 12000 | 0.39 \pm 0.002 | 0.49 \pm 0.001 |
| 5000 | 12000 | 0.51 \pm 0.002 | 0.57 \pm 0.002 |
| 10000 | 12000 | 0.64 \pm 0.001 | 0.73 \pm 0.001 |
| <i>Fixed LRL Size, Varying HRL Size</i> | | | |
| 100 | 12000 | 0.29 \pm 0.001 | 0.40 \pm 0.003 |
| 100 | 5000 | 0.18 \pm 0.002 | 0.34 \pm 0.002 |
| 100 | 1000 | 0.07 \pm 0.025 | 0.20 \pm 0.034 |
| 100 | 500 | 0.03 \pm 0.022 | 0.07 \pm 0.031 |

Table 8: NER performance comparison based on Macro-F1 between AdaMergeX and our approach (BhashaSetu) with English as high-resource and Marathi as low-resource language under varying dataset sizes.

C Detailed Ablation Studies

C.1 Effect of Batch Size on Training Efficiency

To understand the relationship between batch size and training efficiency, we trained GETR-GAT models using batch sizes of 60, 120, and 4096 on a p5en.48xlarge instance. Each batch of size B contained all 100 low-resource language instances

| HRL | Method | Batch Size | Training Steps | Sentiment Classification |
|---------|----------|------------|----------------|--------------------------|
| | | | | Marathi as LRL |
| English | GETR-GAT | 60 | 7141 | 0.73 ± 0.001 |
| | GETR-GAT | 120 | 5416 | 0.73 ± 0.001 |
| | GETR-GAT | 4096 | 4610 | 0.73 ± 0.002 |
| Hindi | GETR-GAT | 60 | 7002 | 0.85 ± 0.001 |
| | GETR-GAT | 120 | 5319 | 0.85 ± 0.001 |
| | GETR-GAT | 4096 | 4220 | 0.85 ± 0.003 |

Table 9: GETR-GAT performance (Macro-F1 score) on Sentiment Classification with Marathi as low-resource language across different batch sizes.

| HRL | Method | Batch Size | Training Steps | Sentiment Classification |
|---------|----------|------------|----------------|--------------------------|
| | | | | Bangla as LRL |
| English | GETR-GAT | 60 | 6981 | 0.72 ± 0.001 |
| | GETR-GAT | 120 | 5501 | 0.72 ± 0.001 |
| | GETR-GAT | 4096 | 4690 | 0.72 ± 0.002 |
| Hindi | GETR-GAT | 60 | 6993 | 0.79 ± 0.001 |
| | GETR-GAT | 120 | 5561 | 0.79 ± 0.002 |
| | GETR-GAT | 4096 | 4399 | 0.79 ± 0.002 |

Table 10: GETR-GAT performance (Macro-F1 score) on Sentiment Classification with Bangla as low-resource language across different batch sizes.

available in the epoch, with the remaining $B-100$ instances sampled from high-resource languages. Of these high-resource samples, 70% were selected based on maximum token overlap with low-resource instances, while the remaining 30% were randomly selected to ensure gradual establishment of cross-lingual connections across epochs.

The motivation for larger batch sizes stems from GETR’s graph-based token connection mechanism. With batch size 4096, nearly all unique tokens from low-resource languages can be connected to high-resource tokens simultaneously, allowing the model to learn global cross-lingual semantics directly. In contrast, smaller batches (e.g., 120) establish these connections incrementally through successive epochs via the 30% random sampling strategy, requiring more steps to achieve saturation.

Our experiments across sentiment classification and NER tasks with Hindi and English as high-resource languages and Marathi and Malayalam as low-resource languages reveal consistent patterns. Training with batch size 4096 achieved comparable performance to batch size 120 while requiring approximately 14% fewer training steps to reach convergence. Batch size 60 required the most steps but maintained similar final performance, confirming that GETR’s effectiveness is robust across different batch sizes. However, due to computational constraints, most practitioners will be limited to smaller batch sizes. Tables 9, 10, 11 and 12 present detailed results across all tasks and language pairs.

| HRL | Method | Batch Size | Training Steps | NER |
|---------|----------|------------|----------------|------------------|
| | | | | Marathi as LRL |
| English | GETR-GAT | 60 | 6981 | 0.40 ± 0.001 |
| | GETR-GAT | 120 | 5501 | 0.40 ± 0.001 |
| | GETR-GAT | 4096 | 4690 | 0.40 ± 0.002 |
| Hindi | GETR-GAT | 60 | 6993 | 0.44 ± 0.001 |
| | GETR-GAT | 120 | 5561 | 0.44 ± 0.001 |
| | GETR-GAT | 4096 | 4399 | 0.44 ± 0.001 |

Table 11: GETR-GAT performance (Macro-F1 score) on NER with Marathi as low-resource language across different batch sizes.

| HRL | Method | Batch Size | Training Steps | NER |
|---------|----------|------------|----------------|------------------|
| | | | | Malayalam as LRL |
| English | GETR-GAT | 60 | 6882 | 0.46 ± 0.001 |
| | GETR-GAT | 120 | 5483 | 0.46 ± 0.001 |
| | GETR-GAT | 4096 | 4552 | 0.46 ± 0.002 |
| Hindi | GETR-GAT | 60 | 7114 | 0.48 ± 0.001 |
| | GETR-GAT | 120 | 5632 | 0.48 ± 0.001 |
| | GETR-GAT | 4096 | 4781 | 0.48 ± 0.003 |

Table 12: GETR-GAT performance (Macro-F1 score) on NER with Malayalam as low-resource language across different batch sizes.

C.2 Robustness to Incomplete Bilingual Lexicons

To evaluate GETR’s resilience to incomplete or noisy cross-lingual edge information, we systematically ablated the percentage of cross-lingual token connections in the graph. In our standard implementation, each batch typically contains 600–1000 token connections established through our translation mappings (using pymultidictionary and manual verification). We conducted experiments retaining 100%, 70%, 50%, 30%, and 0% of these connections through random removal.

The 0% condition serves as a critical baseline: when all cross-lingual edges are removed, GETR-GAT reduces to standard Joint Training, as tokens are only connected within sentences (not across languages). This allows us to quantify the precise contribution of cross-lingual knowledge transfer to overall performance.

The results demonstrate that GETR’s performance degrades gracefully as fewer edges are retained. With 70% edge retention, performance typically drops by only 2–5 percentage points, indicating substantial robustness to incomplete lexicons. Even at 50% retention, models maintain meaningful improvements over Joint Training baselines. The gradual degradation pattern confirms that the approach does not rely on perfect bilingual lexicons and can function effectively with partial

translation information. At 0% edge retention, performance converges precisely to the Joint Training baseline across all tasks and language pairs, validating our hypothesis that cross-lingual edges are the key mechanism enabling GETR’s knowledge transfer capabilities. Table 13 presents comprehensive results across sentiment classification and NER tasks.

C.3 Computational Cost and Environmental Impact

GETR-GAT incurs modest computational overhead compared to Joint Training while delivering substantial performance gains. Training on AWS p4de.24xlarge instances requires approximately 11% additional time per epoch (≈ 50 minutes vs. ≈ 45 minutes for Joint Training) and 8% more peak GPU memory (≈ 41 GB vs. ≈ 38 GB) due to graph neural network computations and neighborhood construction. Over 50 training epochs, the total training time increases from 37.5 hours to 41.7 hours. Inference is approximately 6.3% slower per sample (0.0101 vs. 0.0095 seconds on p3.2xlarge V100 GPUs). The increased energy consumption (≈ 2.2 kWh vs. ≈ 2.0 kWh per full run) results in approximately 10% higher CO₂ emissions (≈ 0.99 kg CO₂ vs. ≈ 0.90 kg CO₂), corresponding to roughly 0.09 kg additional CO₂ per model training. These modest computational costs are justified by the significant performance improvements as shown in Table 1. Detailed report is presented in Table 14.

| HRL | Method | Edge % | Sentiment Classification | | NER | |
|---------|----------------|--------|--------------------------|------------------|------------------|------------------|
| | | | Marathi | Bangla | Marathi | Malayalam |
| English | Joint Training | - | 0.53 \pm 0.002 | 0.63 \pm 0.001 | 0.29 \pm 0.001 | 0.26 \pm 0.002 |
| | AdaMergeX | - | 0.55 \pm 0.001 | 0.63 \pm 0.001 | 0.29 \pm 0.001 | 0.28 \pm 0.002 |
| | GETR-GAT | 100% | 0.73 \pm 0.001 | 0.72 \pm 0.001 | 0.40 \pm 0.001 | 0.46 \pm 0.001 |
| | GETR-GAT | 70% | 0.67 \pm 0.002 | 0.68 \pm 0.001 | 0.38 \pm 0.001 | 0.41 \pm 0.001 |
| | GETR-GAT | 50% | 0.64 \pm 0.001 | 0.66 \pm 0.003 | 0.35 \pm 0.003 | 0.38 \pm 0.001 |
| | GETR-GAT | 30% | 0.58 \pm 0.002 | 0.63 \pm 0.001 | 0.31 \pm 0.001 | 0.32 \pm 0.002 |
| | GETR-GAT | 0% | 0.55 \pm 0.003 | 0.62 \pm 0.001 | 0.28 \pm 0.002 | 0.28 \pm 0.002 |
| Hindi | Joint Training | - | 0.75 \pm 0.004 | 0.67 \pm 0.003 | 0.35 \pm 0.002 | 0.28 \pm 0.002 |
| | AdaMergeX | - | 0.76 \pm 0.001 | 0.69 \pm 0.001 | 0.30 \pm 0.001 | 0.28 \pm 0.001 |
| | GETR-GAT | 100% | 0.85 \pm 0.001 | 0.79 \pm 0.002 | 0.44 \pm 0.001 | 0.48 \pm 0.001 |
| | GETR-GAT | 70% | 0.83 \pm 0.002 | 0.76 \pm 0.002 | 0.40 \pm 0.001 | 0.43 \pm 0.001 |
| | GETR-GAT | 50% | 0.80 \pm 0.001 | 0.72 \pm 0.003 | 0.36 \pm 0.002 | 0.35 \pm 0.001 |
| | GETR-GAT | 30% | 0.77 \pm 0.002 | 0.71 \pm 0.002 | 0.31 \pm 0.001 | 0.32 \pm 0.003 |
| | GETR-GAT | 0% | 0.75 \pm 0.001 | 0.69 \pm 0.001 | 0.30 \pm 0.002 | 0.29 \pm 0.001 |

Table 13: Robustness of GETR-GAT to incomplete bilingual lexicons (Macro-F1 score). Performance is measured when retaining 100%, 70%, 50%, 30%, and 0% of cross-lingual token connections. Sentiment Classification spans Marathi and Bangla low-resource languages, while NER spans Marathi and Malayalam. Average token connections per batch are shown for each edge retention percentage.

| Metric | Joint Training | GETR-GAT |
|--|--|--|
| <i>Training Configuration</i> | | |
| Training Instance | AWS p4de.24xlarge (8 \times A100 80GB) | Same |
| Batch Size | 128 | 120 |
| Number of Epochs | 50 | 50 |
| <i>Training Efficiency</i> | | |
| Training Time per Epoch | \approx 45 min | \approx 50 min (+11%) |
| Total Training Time (50 epochs) | \approx 37.5 hours | \approx 41.7 hours |
| Peak GPU Memory Usage | \approx 38 GB | \approx 41 GB (+8%) |
| <i>Inference Metrics</i> | | |
| Inference Instance | AWS p3.2xlarge (1 \times V100 16GB) | Same |
| Inference Time per Sample | 0.0095 sec | 0.0101 sec (+6.3%) |
| <i>Energy & Environmental Impact</i> | | |
| Energy Consumption per Run | \approx 2.0 kWh | \approx 2.2 kWh (+10%) |
| CO ₂ Emissions per Run | \approx 0.90 kg CO ₂ | \approx 0.99 kg CO ₂ (+10%) |
| <i>Performance Improvements</i> | | |
| Macro-F1 (Sentiment Marathi) | 0.75 | 0.87 (+12 pp) |
| Macro-F1 (Sentiment Bangla) | 0.63 | 0.75 (+12 pp) |
| Macro-F1 (NER Marathi) | 0.35 | 0.44 (+9 pp) |
| Macro-F1 (NER Malayalam) | 0.28 | 0.52 (+24 pp) |

Table 14: Computational cost and environmental impact comparison between Joint Training and GETR-GAT. All training conducted on AWS p4de.24xlarge instances and inference on p3.2xlarge instances. Percentage increases are shown in parentheses. Despite 10–11% overhead in training time and energy consumption, GETR-GAT achieves substantial performance improvements across all tasks.