Revision Summary

July 29, 2025

1 Reviewer: bXRH

Reviewer's comment: In the token embedding transfer approach, there is a strong assumption of a reasonably strong machine translation model from the low resource language to a high resource target language. This is not necessarily true for extremely resource-scarce languages limited this technique's applicability.
Authors' response: As noted in Section 3.3 (L291), our method only requires word-level translations for the small set of unique tokens in the low-resource language training data (approximately 300-400 unique words for our 100-instance datasets). These translations can be created manually without requiring a complete machine translation system, making the approach viable even for extremely low-resource languages. In fact, manual translation may be more reliable than automatic methods for such languages. We have detailed our combined approach using dictionary lookup (when available) and manual verification in L294-L303, highlighting the practicality of this method for genuine low-resource scenarios.

Reviewer's comment: Regarding inference with the graph-enhanced token representation approach, the authors note "During inference, we apply the same principle using training data to form neighborhoods for test instances based on token overlap.". This is an expensive process which might limit realtime application of this technique in low resource languages.
Authors' response: In L600-L605, we have reported the inference time comparison between GETR methods and the Joint Training baseline. Our measurements show that GETR methods increase inference time by only 6.31% compared to the baseline. This minimal overhead makes the approach practical for real-world applications, even in low-resource settings, while providing substantial performance gains (up to 27 percentage points improvement in F1 score).

Reviewer's comment: Apart from joint training and scratch training, the authors could have implemented techniques from data augmentation techniques from the related work section for comparison as baseline methods.
Authors' response: Following this suggestion, we implemented HAL-LRL (Hidden Augmentation Layer within the low-resource language only), which applies mixup-based data augmentation techniques from the related work specifically to the low-resource language data. It is mentioned in L609-L611 and in Table 3.

Reviewer's comment: It would be necessary to add citation(s) to this claim "Traditional ap- proaches of fine-tuning pre-trained models or em- ploying joint training on multilingual architectures often fail to capture the nuanced characteristics of low-resource languages when working with such limited data."
Authors' response: We have modified the statement and added relevant citations to support our claim. The revised text now reads: "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" (L059-L065).

2 Reviewer r74b

Reviewer's comment: The rationale for restricting the datasets to only 100 instances in an extreme low-resource setting, especially when the original datasets contain significantly more data (1K-10K instances), is

unclear and not sufficiently motivated. The authors should explicitly justify this extreme setting, perhaps by highlighting its relevance to real-world scenarios or by providing a comparative analysis with a more moderately low-resource setting.
Authors' response: We have added explicit justification for the extreme setting in L072-L082, highlighting India's linguistic landscape where many languages (such as Dogri, Bodo, Kashmiri, and Santali) have extremely limited digital resources—often fewer than 100 labeled instances for NLP tasks. This represents a genuine real-world challenge. Additionally, as demonstrated in Table 2, we observe that performance degrades dramatically when the low-resource language dataset falls below 100 instances, making this threshold particularly meaningful for studying transfer learning effectiveness. Our chosen setting thus reflects both real-world constraints and a scientifically interesting boundary where traditional approaches begin to fail significantly.

Reviewer's comment: There is no comparison with a baseline using existing cross-lingual transfer methods, such as translate-test (Conneau et al. 2020). They didn't even use a multilingual base model (e.g. XLM-R, instead of monolingual BERTs). Therefore, it is unclear whether the three methods examined are more effective than existing methods.
Authors' response: Our work specifically targets extremely low-resource Indian languages where neither machine translation systems nor pre-trained models exist (as explained in L450-453). In such realistic scenarios, translate-test approaches (Conneau et al. 2020) are not applicable as they require functional NMT systems. However, following this suggestion, we have implemented and evaluated XLM-R fine-tuning under various configurations as a strong baseline. These results are now reported in Tables 1 and 3, showing that XLM-R performs comparably to our Joint Training baseline.

Reviewer's comment: The explanations of the three methods are very difficult to understand. It is unclear whether these three methods are completely new, or applications of existing techniques. For instance, the explanation of hidden layer augmentation is problematic: the cited Zhang 2020 (likely Gong et al. 2022) does not propose the method explained, Chaudhary 2020 is a blog post (which should not be cited as an academic paper), and Feng et al. 2021 is a survey. This method appears to be similar to Mixup (Zhang et al. 2017). Similar issues with clarity and citation exist for the other two methods. The authors should clarify their novelty and provide concise explanations with appropriate citations for each.
Authors' response: Thank you for highlighting these important concerns. We have made significant improvements to address them. We clarify that only the GETR method is claimed as novel in our work (stated explicitly in L116), while HAL and TET are adaptations of existing techniques to the cross-lingual setting. We have corrected the problematic citations for HAL (L233), removing the blog post reference and properly attributing the technique to its original sources, including Mixup (Zhang et al., 2018). For TET, we have provided accurate citations to relevant prior work (L282) that established similar token-level transfer approaches. We have substantially improved the clarity and explanations in sections 3.2, 3.3, and 3.4, making clearer distinctions between our contributions and existing methods, and providing more precise technical descriptions. These revisions ensure better attribution of prior work while highlighting our novel contributions in applying these techniques to extreme low-resource cross-lingual transfer.

Reviewer's comment: The abstract mentions three low-resource languages (Marathi, Bangla, and Malayalam), but only two of them are used in the experiments (Marathi and Malayalam; Table 1 and 2) and inconsistently mentioned as Marahi and Bangla.
Authors' response: We have restructured Table 1 to comprehensively present results across all tasks and all three low-resource languages (Marathi, Bangla, and Malayalam) in the main paper. This ensures consistency between our abstract's mention of these languages (L025) and the experimental results shown in the paper.

Reviewer's comment: The text and numbers in Table 1, 2, and 3 are too small.
Authors' response: We have addressed this formatting issue by restructuring our results into just two tables (Table 1 and 2) in the main section of the paper and significantly increasing the font size of both numbers and text to improve readability.

3 Reviewer Q3aw

Reviewer's comment: The evaluation with respect to Graph Neural Network addition is a bit unfair with the rest of the setups since this method increases the parameter count. A fairer comparison would be to compare against adapters with similar parameter count.
Authors' response: We carefully controlled the parameter count across all architectures to ensure fair comparison, as detailed in L506-515. For GETR methods, we removed transformer layers and added GNN layers to maintain nearly identical parameter counts compared to the baseline models. Additionally, we included adapter fine-tuning (which also has minimal parameter increase) as one of our baseline models in Table 1, providing a direct comparison with another parameter-efficient approach.

Reviewer's comment: The evaluation section felt rushed and disorganised. The setup for the baselines and how the methods from Section 3 are used are introduced while the results are being discussed. These should have been introduced in Section 4.2 and/or in a separate subsection. Section 4.2 is currently a bit confusing since it refers to things introduced later on. Moreover, many of the languages and task setups are not shown in the main text as only results for Marathi on Sentiment Analysis and Malayalam on Named-Entity Recognition are shown, with the rest of the language-task setups shown in the appendix. One potential avenue for showing all results in the main content is to remove the LRL column and have it as part of the column headings, so that you would see different language performances side by side. Similarly for Table 1, it would be more beneficial to see F1 only but multiple languages, instead of seeing accuracy.
Authors' response: We have comprehensively restructured the evaluation section to address these concerns. In the implementation details (Section 4.2), we now clearly introduce all baseline methods before discussing results, ensuring logical flow. We've reorganized Table 1 to include results for all languages and tasks together, showing only macro-F1 scores for consistency and clarity. This comprehensive restructuring eliminates the need to reference the appendix for primary results and provides a clearer side-by-side comparison across language pairs and tasks.

Reviewer's comment: The ACL template states that captions should be below not above the table.
Authors' response: We have corrected this formatting issue by placing all table captions below the tables in accordance with the ACL template requirements.

Reviewer's comment: The tables show standard deviations, for what I can assume is due to multiple runs with different random seeds, but there is no mention of this.
Authors' response: We have addressed this by explicitly mentioning in the caption of Table 1 that all reported results represent the mean and standard deviation across five independent runs with different random seeds. This experimental methodology applies to all results presented throughout the paper.

Reviewer's comment: It would be useful to have some statistical significance testing since some of the numbers seem close to each other.
Authors' response: We have added statistical significance testing through paired t-tests, with p-values reported in Table 1. Details about the significance testing methodology and interpretation are provided in the table caption.

Reviewer's comment: What happens if the Marathi subword is not an actual word? Wouldn't it make more sense to do simple word tokenisation (without subwords) and then subword tokenise after translating?
Authors' response: We agree with this observation. Our implementation already follows this approach: we first translate the complete low-resource language (LRL) word to the high-resource language (HRL), and only afterward tokenize both the original and translated words. We have modified the text in L310-315 to clarify this process and remove any potential confusion about the order of operations.

Reviewer's comment: What translation system is used?
Authors' response: The detailed description of our translation approach for TET is provided in L294-303. As explained there, we use a combination of dictionary lookup (pymultidictionary) followed by manual

verification for accuracy. This is practical for our approach since we only need to translate the limited vocabulary from the low-resource language training data (approximately 300-400 unique words).

Reviewer's comment: How does this work when the script isn't shared? Does this basically mean that the graphs aren't shared between languages?
Authors' response: We address this important question in L417-422, explaining how graph connections are established between languages with different scripts. When languages don't share scripts, we create connections based on translation equivalents rather than direct token matching. The graph remains shared between languages - connections are established between tokens that convey similar meanings across languages, regardless of their script differences, enabling cross-lingual knowledge transfer through the graph structure.

4 Area Chair VEo4

Area Chair's comment: Reviewers express concerns on the baselines used in the paper. The author response clarifies this to some extent, although cross-lingual transfer has been widely studied, so comparisons to more widely used baselines might also be warranted. Further, reviewers suggest significant rewriting / restructuring to improve clarity of the paper.
Authors' response: We appreciate this feedback and have made substantial improvements to address these concerns. We have added comprehensive comparisons with widely used cross-lingual transfer baselines including XLM-R (with four different fine-tuning strategies), adapter fine-tuning, LoRA, and AdaMergeX. These results are now presented in a restructured Table 1, providing a thorough comparison across all language pairs and tasks. Additionally, we have significantly reorganized the paper's structure, particularly in the methodology and evaluation sections, to improve clarity and flow. The implementation details have been consolidated, baselines are now clearly introduced before results are discussed, and the presentation of experimental outcomes has been standardized for easier interpretation.