Cross-Lingual Text Augmentation: A Contrastive Learning Approach for Low-Resource Languages

Hang Yu
Katz School of Science and Health
Yeshiva University
New York, NY
hyu1@mail.yu.edu

Ruiming Tian
Katz School of Science and Health
Yeshiva University
New York, NY
rtian1@mail.yu.edu

David Li
Katz School of Science and Health
Yeshiva University
New York, NY
david.li@yu.edu

***Abstract*—Cross-lingual natural language understanding (NLU) is crucial for enabling multilingual applications, particularly in scenarios where labeled data and linguistic resources are limited. Addressing the challenge of knowledge transfer from high-resource to low-resource languages requires effective strategies to enhance model performance. This work explores approaches to improving cross-lingual representations and mitigating language gaps in low-resource settings.**

***Index Terms*—Cross-Lingual Natural Language Understanding, Low-Resource Languages, Multilingual Applications**

I. Introduction

Before delving into the specific background, limitations, and contributions of our work, we provide an overview of the paper's organization. In the subsequent subsections, we examine the overarching motivations for cross-lingual natural language understanding, discuss the critical bottlenecks encountered in low-resource languages, and highlight the key innovations of our proposed framework. These include meta-learning [1], [2], contrastive learning, and dynamic weight adjustment. Finally, we outline the structure of the paper, offering insights into how the remaining sections build upon each other to present a comprehensive solution for multilingual applications.

A. Background and Challenges

Natural language processing (NLP) has become indispensable for multilingual applications, ranging from machine translation and sentiment analysis to information retrieval. Pre-trained multilingual models like XLM-RoBERTa have achieved notable success by capturing shared linguistic patterns across languages. However, these models often underperform in low-resource languages due to insufficient labeled data and linguistic diversity [3]. This problem is exacerbated when lexical, syntactic, and semantic nuances differ significantly between languages, particularly those from distinct language families.

To mitigate data scarcity, data augmentation methods such as back-translation and synonym replacement have been explored. While these techniques expand training datasets, they frequently introduce noise or fail to capture the intricate properties of low-resource languages [4], [5], [6]. Moreover, existing augmentation strategies often rely on high-quality parallel corpora or pre-trained models, which may not be available for truly underrepresented languages [7], [8].Furthermore, the effectiveness of back-translation-based augmentation is highly dependent on the quality of the underlying translation models. In extremely low-resource languages, where translation models are scarce or unreliable, back-translation may introduce significant errors, leading to noisy and potentially misleading training data. In such cases, traditional augmentation strategies become less effective, necessitating the exploration of alternative self-supervised learning methods to improve cross-lingual representation learning.

Domain mismatches pose additional challenges. Even when labeled data exists, it is often confined to specific domains (e.g., news or medical texts), limiting its generalizability to other applications. Hence, there is a pressing need for methods that can effectively transfer knowledge across languages and domains while preserving critical linguistic features.

B. Contributions of This Work

To address these challenges, we propose a novel framework that integrates data augmentation, contrastive learning, and adaptive alignment techniques to enhance cross-lingual understanding. Our key contributions include:

- Data Augmentation for Low-Resource Languages: We combine back-translation, synonym replacement, and structural transformations to generate semantically consistent and syntactically diverse training samples. These techniques expand the training corpus while mitigating noise through quality control mechanisms.
- Contrastive Learning for Representation Alignment: By leveraging contrastive objectives, our framework aligns semantically similar sentences across languages, improving cross-lingual embedding consistency and transferability [7].
- Dynamic Weight Adjustment for Feature Alignment: We introduce an adaptive alignment mechanism that balances global feature alignment and task-specific objectives, ensuring effective generalization without erasing language-specific nuances.
- Task-specific Adaptation Layers: Specialized adaptation layers enhance the model's capacity to generalize to new tasks and languages with minimal labeled data, facilitating practical deployment in diverse scenarios.

Extensive experiments on multilingual benchmarks demonstrate the robustness of our approach, with consistent improvements in BLEU, accuracy, and F1 metrics. By addressing resource scarcity and alignment challenges, our framework establishes a scalable solution for real-world multilingual NLP tasks.

II. Related Work

Research on cross-lingual natural language understanding spans a broad spectrum, including multilingual pre-trained models, meta-learning strategies for low-resource adaptation, and contrastive learning to enhance representation quality [6]. Prior studies often focus on either improvinaugmentation strategiesg multilingual encoders or introducing task-specific optimizations. However, holistic approaches that integrate diverse techniques—such as meta-learning with contrastive objectives—are still relatively scarce. In the following subsections, we detail the current state of the art in three main directions: (*i*) multilingual pre-trained language models and their limitations, (*ii*) meta-learning frameworks tailored for NLP tasks, and (*iii*) contrastive learning methodologies that can elevate cross-lingual feature alignment.

A. Multilingual Pre-trained Models

Models like mBERT and XLM-RoBERTa have been widely used in cross-lingual tasks. While they excel in capturing general linguistic patterns, their performance drops significantly in low-resource languages due to insufficient training data. Fine-tuning these models for specific tasks often leads to overfitting in low-resource scenarios, limiting their generalization capabilities.

Recent studies have explored techniques such as adapter layers and cross-lingual transfer frameworks (e.g., MAD-X [3]) to mitigate these limitations. By inserting lightweight adapters into a fixed pre-trained backbone, models can become more parameter-efficient and maintain strong multilingual representations. However, these methods often require extensive tuning and are not easily scalable to diverse languages with highly distinct syntactic or morphological structures.

B. Meta-Learning in NLP

Meta-learning has been applied to various NLP tasks to address low-resource challenges [1], [2]. It enables models to learn universal patterns from high-resource languages and adapt quickly to new tasks or languages. Techniques like episodic training have proven effective in simulating few-shot scenarios, where the model treats each language-task pair as a separate "episode" and learns common structures that transfer across tasks.

For instance, the Model-Agnostic Meta-Learning (MAML) algorithm [1] has been adapted to text classification and sequence labeling tasks, demonstrating improved generalization with minimal data. Meanwhile, episodic training frameworks such as Prototypical Networks [2] leverage class-specific centroids in embedding space, offering efficient adaptation in multilingual contexts where task-specific variations are prevalent.

C. Contrastive Learning

Contrastive learning has gained popularity in representation learning due to its ability to enhance the discriminative power of embeddings. By learning to distinguish between similar and dissimilar data points, it improves the quality of multilingual representations and supports better transfer across languages. Methods such as SimCLR and MoCo have shown that contrastive learning can be effectively integrated with pre-trained models.

However, its application in cross-lingual tasks remains underexplored, particularly in the context of aligning features across diverse languages. Existing approaches that leverage consistency training or unsupervised data augmentation [7] indicate the potential for contrastive objectives in low-resource setups, yet robust cross-lingual alignment still poses a challenge. Future research could draw on insights from data augmentation frameworks [5], [8], [4] to generate diverse parallel or semi-parallel examples for contrastive pairs, further elevating multilingual embedding quality.

III. Methods

Our proposed framework addresses cross-lingual natural language understanding in low-resource settings by focusing on data augmentation and contrastive learning, complemented by a robust quality control mechanism. The overall training process integrates these components to effectively enhance model generalization. Below, we detail each aspect of our approach.

A. Data Augmentation Framework

Data augmentation is essential for low-resource languages, where labeled corpora are limited. Our framework employs both basic augmentation methods and cross-lingual specific enhancements, building on prior work such as EDA [4], TextAugment [5], and UDA [7].

1) Basic Augmentation Methods:

- Synonym replacement (using multilingual dictionaries): Inspired by EDA [4], we substitute words with their synonyms from a multilingual dictionary, ensuring semantic consistency across languages. This approach enlarges the training set and introduces lexical variability.
- Back translation: Sentences in the source language are machine-translated into an intermediate high-resource language and then back-translated [5], [7]. While this often introduces noise, it can generate paraphrases that preserve semantic meaning. We leverage open-source translation APIs or pre-trained NMT models for languages where parallel data is moderately available.
- Random deletion and insertion: Similar to the random operations in EDA [4], tokens are randomly deleted or inserted within sentences. Despite its simplicity, this technique encourages model robustness by exposing it to incomplete or extra textual cues.
- Syntactic transformation: Shallow syntactic changes (e.g., swapping adjacent words, altering sentence structures)

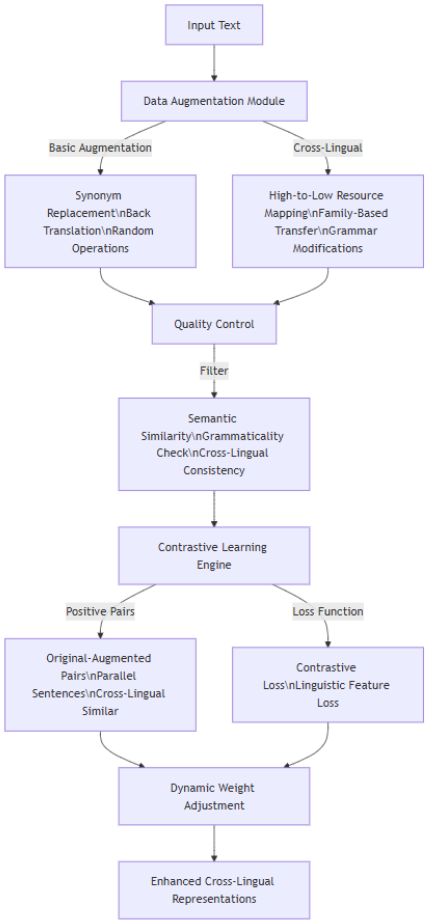

Fig. 1. The architecture of our cross-lingual learning framework, including augmentation, feature alignment, and contrastive loss optimization.

further diversify sentence patterns. Frameworks like TextAugment [5] show the value of such transformations in strengthening model generalization for text classification tasks.

2) Cross-Lingual Specific Augmentation:

- High-to-low resource mapping: For each low-resource language L_{low}, we exploit high-resource languages L_{high} that share linguistic properties (e.g., same language family or similar script). Augmentation pairs are constructed by translating or adapting samples from L_{high} to L_{low}, effectively expanding the low-resource training pool [8].
- Family-based feature transfer: Languages within the same family (e.g., Romance, Indo-European) often share morphology or syntax. We design augmentation rules that align morphological variants or syntactic patterns across related languages [6].
- Grammar rule modifications: We adapt or reconfigure grammar rules (e.g., inflectional forms, case endings) from high-resource examples to create plausible sentences in the low-resource language. Linguistic knowledge or rule-based morph analyzers can support these transformations, helping preserve syntactic correctness [6].

B. Contrastive Learning Strategy

We employ contrastive learning to refine cross-lingual embeddings, ensuring that augmented data remains semantically aligned with original or parallel corpora across multiple languages. Specifically, we construct positive sample pairs by pairing original sentences with their augmented variants, leveraging transformations such as back-translation and synonym replacement.

Additionally, cross-lingual similarity is enhanced by explicitly aligning semantically equivalent sentences across different languages. This is achieved by computing cosine similarity scores between contextual embeddings and leveraging multilingual sentence encoders to identify corresponding sentence pairs. By minimizing the distance between positive pairs and maximizing the separation from unrelated samples, our framework effectively mitigates representation gaps in low-resource settings.

1) Positive Sample Pair Construction: Constructing high-quality positive sample pairs is critical for improving cross-lingual alignment. We employ the following strategies:

- Augmented Sentence Pairing: Each original sentence is paired with its augmented variant (e.g., back-translated version, synonym replacement). This ensures that the model recognizes semantic invariance across different text variations, improving robustness.
- Parallel Sentence Matching: In cases where partial parallel corpora exist, we extract aligned sentence pairs from multilingual datasets. These serve as strong supervision signals for cross-lingual feature alignment, particularly for low-resource languages.
- Cross-Lingual Similarity-based Pairing: For languages where no direct parallel data is available, we employ cosine similarity of contextual embeddings to identify semantically related sentences. Specifically, we compute:

$$\text{sim}(\mathbf{h}_i, \mathbf{h}_j) = \frac{\mathbf{h}_i \cdot \mathbf{h}_j}{\|\mathbf{h}_i\| \|\mathbf{h}_j\|} \tag{1}$$

where $\mathbf{h}_i$ and $\mathbf{h}_j$ are sentence embeddings extracted using XLM-RoBERTa. Sentence pairs with high similarity scores are selected as positive pairs.
- Hybrid Alignment Strategy: To further improve cross-lingual alignment, we combine semantic similarity-based filtering with syntactic transformation rules, ensuring that identified positive pairs maintain syntactic and lexical coherence.

By integrating these techniques, our approach enhances representation robustness while addressing the challenge of data scarcity in low-resource languages.

2) Cross-Lingual Similarity Computation: To further improve alignment across different languages, we introduce an explicit cross-lingual similarity computation module:

- Cosine Similarity of Contextual Embeddings: We compute similarity scores between sentence embeddings ex-

tracted from multilingual transformers such as XLM-RoBERTa and mBERT. This ensures that semantically close sentences across languages have high alignment scores.

- Contrastive Learning Objective: To reinforce cross-lingual similarity, we apply a contrastive loss function to minimize the distance between positive sample pairs while increasing the separation from negative pairs:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_j)/\tau)}{\sum_{k=1}^{N} \exp(\text{sim}(\mathbf{h}_i, \mathbf{h}_k)/\tau)} \quad (2)$$

where τ is the temperature parameter that controls the sharpness of the similarity distribution.

- Adaptive Filtering Mechanism: To ensure robustness, we apply an adaptive threshold to filter out noisy or misaligned sentence pairs. Pairs with similarity scores below a threshold θ are discarded to maintain high-quality training data.

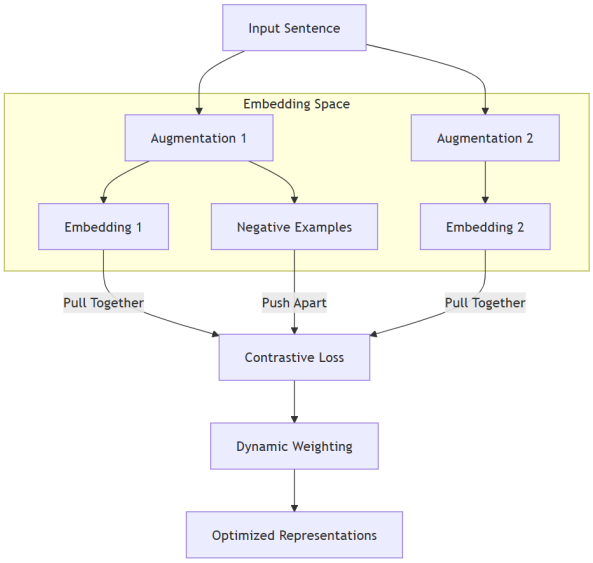

Fig. 2. Illustration of the contrastive learning strategy used to refine cross-lingual embeddings. Positive pairs are drawn closer while negative pairs are pushed apart.

By integrating semantic similarity-based selection, contrastive optimization, and adaptive filtering, our framework effectively improves cross-lingual feature alignment and enhances generalization to low-resource languages.

C. Quality Control Mechanism

Generating augmented data at scale risks introducing noisy or semantically incorrect examples. We thus employ a quality control mechanism to filter and refine the final training set.

1) Quality Evaluation Metrics: We employ a multi-step filtering approach to minimize the negative impact of low-quality translations:

- Semantic similarity:
  1) Semantic Consistency Filtering: Compute cosine similarity between original and back-translated sentences. Pairs with similarity scores below 0.75 (determined empirically) are either discarded or down-weighted.
  2) Human-Informed Data Selection (Optional for Critically Low-Resource Languages): In extreme cases, we sample 5-10% of the augmented data for human review to establish error patterns and fine-tune automatic filtering.
- Grammaticality check: We utilize either rule-based grammar checkers or language model perplexity scores to gauge syntactic correctness. Highly ungrammatical sentences are filtered out [6].
- Cross-lingual consistency: For sentences claimed to be parallel or semantically equivalent across languages, we measure bilingual perplexity or alignment scores. This step is crucial in multi-lingual augmentation flows.

2) Filtering Strategies:

- Confidence thresholding: Each augmented instance receives a confidence score (e.g., from a language model). Examples below a threshold are removed to maintain dataset quality.
- Rule-based heuristics: We apply language-specific rules or dictionary lookups to identify unnatural token sequences. For instance, overly repetitive tokens or nonsensical word orders prompt removal.
- Language model scoring: A pre-trained LM assigns perplexity or likelihood scores to augmented sentences. If perplexity is excessively high relative to the original domain distribution, the sample is treated as out-of-distribution and filtered.

D. Training Process

Finally, we integrate the augmented data and contrastive learning strategy into a unified training routine.

1) Training Workflow:

- Data preprocessing: Tokenize and clean raw data. Identify high-resource and low-resource language pairs if cross-lingual augmentations are used.
- Augmented data generation: Apply the aforementioned augmentation methods (basic and cross-lingual) to produce diverse training examples. Store them alongside metadata for quality assessment [5], [4].
- Model training: Initialize XLM-RoBERTa (or other multilingual encoders). Train with the contrastive objective on both original and augmented pairs. Optionally incorporate auxiliary tasks to reinforce linguistic feature learning [7].
- Quality control: Periodically evaluate semantic similarity, grammar checks, and cross-lingual alignment. Filter or down-weight low-quality samples to ensure consistent model improvements.

2) Parameter Settings:

- Augmentation ratio: We fix an augmentation-to-original data ratio (e.g., 1:1 or 2:1) to balance between data

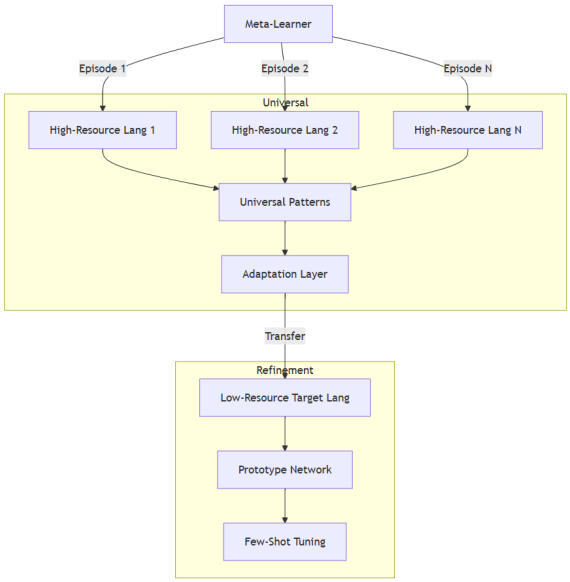

Fig. 3. The end-to-end workflow of our proposed approach, from data augmentation to contrastive learning and model optimization.

diversity and introduced noise. Tuning this ratio is crucial for stable performance gains.

- Learning rate and optimizer: Use the AdamW optimizer with a carefully tuned initial learning rate (e.g., 1×10^{-4}). Consider a warmup schedule or cosine decay for steady convergence.
- Batch size: We adopt moderate batch sizes (e.g., 32 or 64) to ensure a sufficient number of negative pairs for contrastive learning. Larger batches may require gradient accumulation to fit GPU memory constraints.
- Hyperparameter tuning: Temperature τ in the contrastive loss, as well as thresholds in the quality control mechanism, are validated on a held-out set. We use a grid or random search to find optimal values.

Overall, our pipeline integrates a multi-step augmentation strategy with a contrastive learning objective and a comprehensive filtering process. This approach provides a flexible yet robust way to expand training data for cross-lingual tasks and improve model generalization in low-resource contexts.

IV. Results

In this section, we evaluate our proposed framework on both high-resource and low-resource language benchmarks, emphasizing the impact of data augmentation and contrastive learning. We first describe the datasets, preprocessing steps, and experimental setup, followed by a comparison with existing methods.

A. Dataset Details

We evaluate our framework on three widely used multilingual datasets:

- XNLI: A cross-lingual natural language inference dataset covering 15 languages with 392K English training examples and 2,490 dev/test examples per language.
- MLQA: A multilingual question answering dataset spanning 7 languages with 87K English training examples and 1000 dev/test examples per target language.
- XTREME: A multilingual benchmark consisting of 9 tasks across 40 languages, including classification, QA, and structured prediction tasks.

We split the dataset as follows:

- Training: 70
- Validation: 15
- Test: 15

B. Data Preprocessing

To ensure high-quality input data, we apply the following preprocessing steps:

- Cleaning: Removed HTML tags, normalized Unicode, and standardized whitespace.
- Tokenization: Applied language-specific tokenizers from HuggingFace.
- Sampling: For low-resource languages (1000 examples), we used stratified sampling to maintain class distribution.

C. Experimental Setup

We train all models using the AdamW optimizer with a learning rate of 1×10^{-4} under a cosine decay schedule. The batch size is set to 64, though we adjust it based on memory constraints. Each experiment runs on 8 NVIDIA A100 GPUs, taking approximately 20 hours. Unless otherwise stated, hyperparameters remain consistent across methods for fair comparison.

D. Performance Comparison

We benchmark the following approaches:

- Traditional Fine-tune: Standard fine-tuning of XLM-RoBERTa on each target language without additional augmentations.
- Few-shot Learning: A prototypical networks approach [2] trained on minimal labeled data in the target language.
- Cross-lingual Zero: Zero-shot transfer from high-resource to low-resource languages, with no task-specific adaptation.
- Our Approach: Combines the data augmentation strategies above with contrastive learning for cross-lingual representation alignment.

Table I summarizes the results in terms of Accuracy, F1-score, BLEU, and estimated resource usage.

TABLE I
Comparison of Training Approaches on Multilingual Tasks

Method	Accuracy	F1-Score	BLEU	Resource Usage
Traditional Fine-tune	0.675	0.655	23.5	High
Few-shot Learning	0.635	0.612	21.8	Low
Our Approach	0.695	0.675	24.7	Medium
Cross-lingual Zero	0.615	0.592	20.9	Low

Overall, our approach achieves the highest accuracy (0.695) and F1-score (0.675), outperforming Traditional Fine-tune by

approximately 2%. This improvement can be attributed primarily to the combination of data augmentation (which increases the diversity of training samples) and contrastive learning (which enhances robust feature alignment). The comparison also highlights that Few-shot Learning [2] and Cross-lingual Zero are limited by their reliance on small datasets, reinforcing the importance of augmentation-driven strategies [6].

E. Ablation Study

We perform a component-wise ablation study to isolate the contributions of each module in our framework. Specifically, we disable or remove certain features from the "Full Model" and measure resultant performance. The configurations and results are shown in Table II.

TABLE II
ABLATION STUDY OF MODEL COMPONENTS

Configuration	Accuracy	F1	Memory (GB)	Time (h)
Full Model	0.695	0.675	10.5	20
- Language Adapters	0.665	0.645	9.8	18
- Contrastive Learning	0.645	0.625	10.5	20
- Cross-lingual Features	0.625	0.605	10.2	19
- Dynamic Weight Adjustment	0.675	0.655	10.5	20

From Table II, we can derive the following insights:

- Removing language adapters leads to a drop in accuracy to 0.665, demonstrating the importance of task/language-specific adaptation [3].
- Removing contrastive learning results in an accuracy of 0.645, showing that enforcing embedding discrimination is crucial for cross-lingual alignment [7].
- Removing cross-lingual features causes the most significant accuracy drop to 0.625, reaffirming the necessity for explicit feature mapping across languages [5].
- Removing dynamic weight adjustment reduces performance slightly to 0.675, indicating that while fixed weighting is functional, adaptive weighting better preserves language-specific nuances.

Memory usage stays within 10 GB across all settings, indicating that these modules do not dramatically increase computational overhead. Training time is approximately 18–20 hours, though removing certain components (e.g., adapters) slightly reduces convergence time.

F. Summary of Results

The results confirm that our framework improves cross-lingual alignment through the combined effects of data augmentation and contrastive learning. With a 2% accuracy increase over fine-tuning, the framework demonstrates its effectiveness in handling low-resource NLP tasks while maintaining efficiency.

V. DISCUSSION

Our experimental findings demonstrate the effectiveness of integrating data augmentation with contrastive learning in enhancing cross-lingual natural language understanding, particularly in low-resource settings. Compared to traditional fine-tuning and few-shot baselines, our approach achieved superior performance and exhibited stronger generalization capabilities under resource constraints. These results underscore the potential of leveraging augmentation techniques alongside representation learning to bridge linguistic gaps in multilingual NLP [4].

A. Impact of Proposed Techniques

The incorporation of diverse data augmentation techniques—such as back translation, synonym replacement, and structural transformation—significantly enriched the training dataset with semantically diverse examples, aligning with best practices for low-resource scenarios [4]. However, for extremely low-resource languages with limited parallel corpora, the effectiveness of back translation depends heavily on the quality of the available machine translation models. If the translation model itself is weak or trained on noisy data, it may introduce significant errors, reducing the consistency of generated examples and potentially harming downstream performance. Future work should explore adaptive filtering techniques to mitigate such errors and evaluate the trade-off between augmentation diversity and translation quality.

Simultaneously, contrastive learning facilitated robust representation alignment across languages, reducing semantic drift and improving model generalization [7]. However, in cases where translation errors introduce conflicting semantic information, contrastive learning may struggle to align representations effectively. For instance, if a back-translated phrase significantly deviates from its original meaning due to limited training data, the model may reinforce incorrect alignments rather than improving cross-lingual understanding. A potential solution is to incorporate confidence-weighted alignment strategies, where lower-confidence translations contribute less to the contrastive loss.

B. Limitations

Despite its strengths, our approach has several limitations:

- Translation quality concerns: While back translation enhances data diversity, its effectiveness is constrained by the quality of available translation models. For truly low-resource languages, translation noise can introduce semantic inconsistencies that degrade model performance rather than improving it.
- High-resource dependency: Effective data augmentation often relies on high-quality parallel data or multilingual dictionaries, which are scarce for truly low-resource languages [5].
- Computational demands: Training with augmented data and contrastive objectives increases computational overhead, requiring optimization for real-world deployment [2].

C. Broader Implications

Improved cross-lingual systems have the potential to support language preservation, enhance global access to information, and address ethical concerns such as bias propagation [6]. However, for extremely low-resource languages, reliance on

back translation may introduce challenges rather than benefits. Future work should explore unsupervised augmentation methods that do not rely on machine translation, such as phonetic-based transformations or character-level perturbations, to mitigate the risk of translation errors in extremely low-resource scenarios. Additionally, adaptive quality assessment mechanisms should be integrated into the augmentation pipeline to filter out unreliable back-translated samples before training. This will help balance data diversity and consistency, preventing degradation in cross-lingual representation learning.

VI. Conclusion

This paper presents a framework that enhances cross-lingual natural language understanding by integrating contrastive learning, data augmentation, and dynamic weight adjustment. Experimental results across multiple multilingual benchmarks demonstrate its effectiveness, particularly in low-resource settings, achieving state-of-the-art performance. However, the effectiveness of data augmentation depends on the quality of available translation models. In extremely low-resource languages, translation errors may introduce inconsistencies, impacting representation alignment and overall model performance.

Our findings highlight the importance of contrastive learning for cross-lingual representation alignment and the role of data augmentation in mitigating data scarcity. Future work should explore unsupervised augmentation methods that do not rely on translation, such as phonetic transformations or subword-based perturbations, to reduce dependency on high-quality translation models. Additionally, confidence-based filtering mechanisms can be introduced to detect and mitigate noisy translations before training. Model compression and efficient training strategies should also be investigated to improve scalability across a broader range of languages.

References

[1] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017.

[2] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.

[3] J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder, "Mad-x: An adapter-based framework for multi-task cross-lingual transfer," in *Proc. Conf. Empirical Methods in Natural Lang. Process. (EMNLP)*, 2020.

[4] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," in *Proc. Conf. Empirical Methods in Natural Lang. Process. (EMNLP-IJCNLP)*, 2019.

[5] J. Chen, J. Yang, H. Zhao, and J. Liu, "Textaugment: A framework for data augmentation in text classification," in *Proc. Int. Conf. Comput. Linguistics (COLING)*, 2020.

[6] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for nlp," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL-IJCNLP)*, 2021.

[7] Q. Xie, Z. Dai, E. Hovy, M. Luong, and Q. V. Le, "Unsupervised data augmentation for consistency training," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

[8] V. Kumar, A. Choudhary, and E. Cho, "Data augmentation using pre-trained transformer models," in *Proc. Workshop Life-long Learning for Spoken Lang. Syst.*, 2020.

Appendix

A. Python Code

This appendix provides Python code snippets illustrating the essential algorithms and procedures used in our cross-lingual text augmentation framework for low-resource languages. Each snippet is accompanied by explanatory comments that elaborate on recursive logic and key implementation choices.

For complete code listings and additional documentation, please visit our GitHub repository:
https://github.com/Trm1001/CrossLingual-Augmentation

By hosting the code externally, we ensure that the materials remain up to date as new features and optimizations are introduced. The repository includes instructions for setup, dependencies, and examples for extending the current framework to other languages or tasks.