NeighXLM: Enhancing Cross-Lingual Transfer in Low-Resource Languages via Neighbor-Augmented Contrastive Pretraining

Anonymous ACL submission

Abstract

Recent progress in multilingual pretraining has yielded strong performance on high-resource languages, albeit with limited generalization to genuinely low-resource settings. While prior approaches have attempted to enhance cross-lingual transfer through representation alignment or contrastive learning, they remain constrained by the extremely limited availability of parallel data to provide positive supervision in target languages. In this work, we introduce NeighXLM, a neighbor-augmented contrastive pretraining framework that enriches target-language supervision by mining semantic neighbors from unlabeled corpora. Without relying on human annotations or translation systems, NeighXLM exploits intra-language semantic relationships captured during pretraining to construct high-quality positive pairs. The approach is model-agnostic and can be seamlessly integrated into existing multilingual pipelines. Experiments on Swahili demonstrate the effectiveness of NeighXLM in improving cross-lingual retrieval and zero-shot transfer performance.

1 Introduction

Recent progress in natural language processing (NLP) has brought impressive performance to English and other high-resource languages across a wide range of tasks. However, for genuinely low-resource languages, models still struggle due to the lack of labeled data and effective transfer. Early multilingual models such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), and XLM-R (Conneau et al., 2020) exhibit some cross-lingual transfer capabilities, but their alignment remains limited, especially for typologically distant languages.

To mitigate this limitation, recent research has focused on incorporating explicit cross-lingual signals into pretraining at multiple levels of granularity, including token-level (Luo et al., 2021; Zhang et al., 2023), word-level (Huang et al., 2019; Cao et al., 2020; Ji et al., 2021), sentence-level (Chi et al., 2021; Ouyang et al., 2021), and syntax-level (Wu and Lu, 2023; Ahmad et al., 2021; He et al., 2019). These methods enhance alignment by modeling cross-lingual consistency at their respective levels, providing stronger supervision across languages. In addition, contrastive learning techniques have shown strong potential in improving sentence representations, both in monolingual (e.g., SimCSE; Gao et al. 2021) and multilingual (e.g., ConSERT; Yan et al. 2021 and LaBSE; Feng et al. 2022) contexts.

Despite recent advances such as alignment-based techniques and contrastive pretraining that better exploit existing corpora, a fundamental bottleneck persists: the scarcity of labeled data for low-resource languages. To address this, recent work explores pseudo-supervision strategies that simulate labeled pairs from monolingual corpora. ERNIE-M (Ouyang et al., 2021), for example, constructs pseudo-parallel sentence pairs via back-translation, but the resulting supervision is only as reliable as the underlying translation system, which often generates noisy or semantically inaccurate outputs in low-resource settings due to the scarcity of parallel training data. Alternatively, Keung et al. (2020) mine cross-lingual neighbors in embedding space as training pairs. However, in the absence of strong initial alignment, particularly for typologically distant and under-resourced language pairs, cross-lingual nearest neighbors in the embedding space may not be semantically aligned. Training on such misleading neighbors can reinforce incorrect associations and degrade cross-lingual generalization. This highlights a core challenge: how to obtain more high-quality labeled supervision for low-resource languages.

In this paper, we propose NeighXLM, a neighbor-augmented contrastive pretraining framework that enriches target-language supervision

without relying on translation systems. While annotated data are scarce, large unlabeled corpora are often available, even for low-resource languages. Pretrained multilingual encoders, trained on language modeling objectives, implicitly capture intra-language semantic relationships by positioning semantically similar sentences closer in the embedding space. NeighXLM exploits this property by mining semantically similar neighbors from unlabeled corpora, thereby constructing high-quality positive pairs to enhance contrastive pretraining. Figure 1 illustrates the NeighXLM framework. We evaluate NeighXLM on Swahili (sw), covering diverse downstream tasks including cross-lingual sentence retrieval and zero-shot transfer tasks such as classification and question answering. Results across multiple benchmarks show that NeighXLM consistently outperforms the base model, demonstrating its effectiveness in enhancing cross-lingual transfer for genuinely low-resource languages.

2 Related Work

2.1 Multilingual Pretraining and Cross-Lingual Alignment

Early multilingual models, such as mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019) and XLM-R (Conneau et al., 2020), demonstrated that even simple pretraining objectives, like multilingual masked language modeling (MMLM) and translation language modeling (TLM), could endow models with non-trivial cross-lingual transfer abilities. However, multilingual representations often cluster sentences by language rather than meaning due to insufficient cross-lingual alignment (Libovický et al., 2020), and substantial transfer gaps persist for genuinely low-resource languages (Wu and Dredze, 2020).

To address representational misalignment, recent research has focused on injecting explicit cross-lingual signals into pretraining objectives at various linguistic levels. Token-level methods such as VECO (Luo et al., 2021) and VECO 2.0 (Zhang et al., 2023) enhance cross-lingual alignment by introducing a plug-in cross-attention module into masked token prediction tasks or by directly applying contrastive loss to aligned token pairs. Word-level methods like Unicoder (Huang et al., 2019), Word-aligned BERT (Cao et al., 2020), and word reordering (Ji et al., 2021) focus on the importance of words, aligning them across languages by targeting word pairs or addressing cross-lingual differences in word order. Syntax-aware methods—such as StructXLM (Wu and Lu, 2023), Syntax-augmented BERT (Ahmad et al., 2021), and projection-based approach (He et al., 2019)—enhance cross-lingual transfer by integrating syntactic structures, either through explicit syntactic annotations or unsupervised discovery, into training objectives; typology-guided methods (Ji et al., 2023) further supplement this by incorporating language-level features such as canonical word order (e.g., SVO vs. SOV). At the sentence level, models such as InfoXLM (Chi et al., 2021) and ERNIE-M (Ouyang et al., 2021), along with many of the aforementioned approaches, employ translation ranking or contrastive learning objectives to align cross-lingual sentence representations.

2.2 Contrastive Learning for Sentence Representations

Contrastive learning has emerged as a powerful tool for learning semantically meaningful representations. Early vision models like SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) inspired sentence-level approaches in NLP. SimCSE (Gao et al., 2021) uses dropout-based augmentation for unsupervised contrastive learning, and NLI entailment pairs for the supervised variant. ConSERT (Yan et al., 2021) applies semantic-preserving data augmentations—such as token shuffling, cutoff, and adversarial dropout—to construct contrastive pairs. In the multilingual settings, LaBSE (Feng et al., 2022) aligns cross-lingual sentence representations using translation pairs as positives in a dual-encoder setup, and mSimCSE (Wang et al., 2022) extends the SimCSE framework to multilingual settings.

2.3 Pseudo-supervision and Neighbor Mining

Despite the advances achieved by alignment-based and contrastive learning techniques, low-resource languages still suffer from limited high-quality supervision, motivating alternative enhancement strategies. A common approach is to mine pseudo-positive pairs from monolingual corpora, thereby simulating supervision without human annotation. For example, ERNIE-M (Ouyang et al., 2021) employs back-translation to generate synthetic sentence pairs; however, the quality of this supervision is highly dependent on the accuracy of the translation model, which itself depends on the availability of parallel corpora—a resource often absent in low-resource settings. This creates a vicious cycle: poor

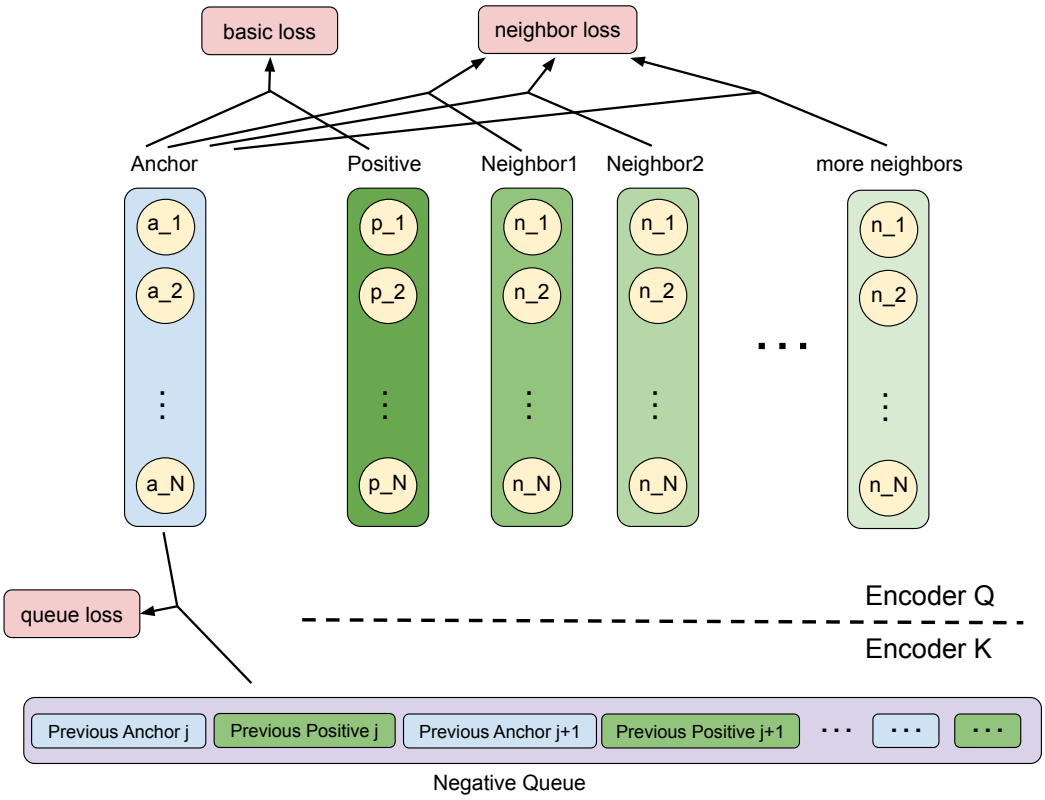

Figure 1: Overview of **NeighXLM**. Given a batch of source–target sentence pairs (anchor-positive), NeighXLM augments each positive with k semantic neighbors mined from unlabeled target-language corpora. The main encoder Q encodes anchors, positives and neighbors for current contrastive learning, while the momentum encoder K encodes previous samples to populate a dynamic negative queue.

translations weaken supervision and hinder cross-lingual alignment. Keung et al. (2020) propose to mine cross-lingual sentence pairs from unlabeled corpora by treating nearest neighbors in embedding space as positives. While effective in some cases, this approach may suffer in the context of linguistically distant and low-resource language pairs (e.g., Swahili–English, which differ substantially in syntax, morphology, and script), where the initial cross-lingual embedding neighborhoods may be noisy or misaligned. Training on such unreliable alignments risks amplifying semantic inconsistencies rather than correcting them.

3 Method

3.1 Overview

In this paper, we propose NeighXLM, a neighbor-augmented contrastive pretraining framework that enriches target-language supervision by mining semantically similar neighbors from unlabeled corpora. The overall framework of NeighXLM is illustrated in Figure 1. We assume access to a small set of source-to-target parallel sentences—typically in the order of a few thousand—which serve as the seed supervision for cross-lingual contrastive learning. In our setup, we refer to the source-language sentence as the *anchor*, and its corresponding target-language translation as the *positive*. Starting from a batch of (anchor, positive) pairs, we retrieve k nearest semantic neighbors for each positive sample from the unlabeled target corpus. We maintain two encoders during training: a main encoder Q that is updated through standard back-propagation, and a momentum encoder K whose parameters are updated as an exponential moving average of Q. For previously seen anchor and positive samples, we encode them using K and store their embeddings into a dynamic queue, serving as a repository of *negative* examples. For the current batch, we encode anchor, positive and neighbor sentences with Q. Contrastive learning is applied to bring anchor embeddings close to their positives and semantic neighbors, while pushing anchors away from negatives stored in the queue.

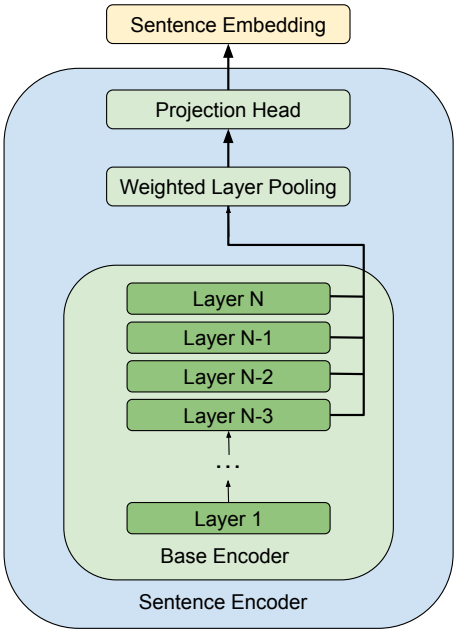

Figure 2: Sentence encoder architecture.

3.2 Neighbor-Augmented Input Construction

We start with a set of parallel source-target sentence pairs. For each target sentence, we retrieve k semantic neighbors from a large unlabeled corpus in the target language. To select these neighbors, we compute semantic similarity based on the cosine distance between sentence embeddings produced by a multilingual encoder, where embeddings are obtained via mean pooling over the hidden states. Thus, the final input to the model includes the anchors (source language), the positives (target language) and multiple semantic neighbors (target language).

3.3 Encoder Architecture

The structure of our sentence encoder is illustrated in Figure 2. The base encoder is a pretrained multilingual model (e.g., InfoXLM or XLM-R). For each input sentence, we extract the last four hidden layers and perform weighted layer pooling (WLP) to produce a rich contextualized representation. Specifically, we learn trainable scalar weights over the selected layers and compute a weighted sum. Following SimCLR (Chen et al., 2020), we add a two-layer projection head with nonlinear activation to map the pooled representation into a contrastive space. Contrastive training is conducted in this projected space, which has been shown to help base encoders yield better downstream task representations.

3.4 Contrastive Learning with Additive Margin

Given a training batch of N anchor–positive pairs, each anchor has one positive sample and treats the remaining $N-1$ samples as negatives. We use cosine similarity as the base similarity function, denoted by $\phi(x, y) = \cos(f(x), f(y))$, where $f(\cdot)$ denotes the output of the sentence encoder. We apply an additive margin (Yang et al., 2019) to the positive logits and incorporate temperature scaling (Chen et al., 2020) directly into the similarity function. The modified similarity is defined as:

$$\tilde{\phi}(x_i, y_j) = \begin{cases} \frac{\phi(x_i,y_j)-m}{\tau}, & \text{if } i = j \\ \frac{\phi(x_i,y_j)}{\tau}, & \text{otherwise} \end{cases} \quad (1)$$

The contrastive loss for the source-to-target direction is:

$$\mathcal{L}_{x\to y} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\tilde{\phi}(x_i,y_i)}}{\sum_{j=1}^{N} e^{\tilde{\phi}(x_i,y_j)}} \quad (2)$$

We adopt a bidirectional contrastive objective:

$$\mathcal{L}_{\text{basic}} = \mathcal{L}_{x\to y} + \mathcal{L}_{y\to x} \quad (3)$$

3.5 Momentum Encoder and Queue Mechanism

To stabilize training with dynamic negatives, we maintain two encoders: a main encoder Q and a momentum encoder K, following MoCo (He et al., 2020). Let θ_q and θ_k denote the parameters of the main encoder and momentum encoder, respectively. After each training batch, the momentum encoder is updated via an exponential moving average:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q$$

where m is a momentum coefficient close to 1. During training, the current batch samples are processed as follows:

- Anchor embeddings h_a and positive embeddings h_p are computed using the main encoder Q.
- Negative embeddings $\{h_{n_1}, h_{n_2}, \ldots, h_{n_{qs}}\}$ are retrieved from the dynamic queue of size qs, where all entries are encoded by the momentum encoder K.

For each anchor embedding h_{a_i}, the similarity logits are constructed as:

$$\text{logits}_i = [\phi(h_{a_i}, h_{p_i}), \phi(h_a, h_{n_1}), \phi(h_a, h_{n_2}), \ldots, \phi(h_a, h_{n_{qs}})] \quad (4)$$

where $\phi(\cdot, \cdot)$ denotes cosine similarity. The first position corresponds to the positive sample, and the remaining positions correspond to negatives. We compute the InfoNCE (Oord et al., 2018) loss by applying a cross-entropy objective over the logits, with the ground-truth label set to 0 (indicating the positive sample). The queue-based contrastive loss for a batch of N anchors is:

$$\mathcal{L}_{\text{queue}} = \frac{1}{N} \sum_{i=1}^{N} \text{CrossEntropy}(\text{logits}_i, 0)$$

where logits_i denotes the logits for the i-th anchor.

After each training step, we use the momentum encoder K to recompute the embeddings of the current batch's anchors and positives, and enqueue them into the memory queue for future negative sampling.

3.6 Neighbor-Augmented Contrastive Objective

To further enrich supervision, NeighXLM leverages semantic neighbors. For each positive sample h_{p_i} inside the batch, k semantic neighbors $\{h_{n_i}^{(1)}, \ldots, h_{n_i}^{(k)}\}$ are sampled. The neighbor-augmented contrastive loss is computed batch-wise:

$$\mathcal{L}_{\text{neighbor}} = \sum_{k=1}^{K} \frac{1}{k} \Big(\mathcal{L}(h_a, h_n^{(k)}) + \mathcal{L}(h_n^{(k)}, h_a) \Big) \quad (5)$$

where $\mathcal{L}(\cdot, \cdot)$ denotes the standard InfoNCE loss without an additive margin. Unlike direct translation pairs, these semantic neighbors are approximate matches mined from unlabeled data and may not guarantee precise semantic equivalence. To prevent over-constraining their representations, we omit the margin term and apply vanilla InfoNCE. The inverse rank-based weighting $\frac{1}{k}$ reflects the intuition that top-ranked neighbors are semantically closer and thus more reliable. This design encourages the model to place greater emphasis on high-quality neighbors while still incorporating broader contextual signals. The reduced weight on lower-ranked neighbors is particularly helpful when the base encoder produces suboptimal representations or the unlabeled corpus is limited in size or diversity—conditions under which lower-ranked neighbors are more likely to be semantically noisy or misaligned. Consequently, the weighting scheme enhances training stability and robustness in challenging low-resource scenarios.

3.7 Overall Training Objective

The overall training loss aggregates the basic contrastive loss, the negative queue contrastive loss, and the neighbor-augmented contrastive loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{basic}} + \mathcal{L}_{\text{queue}} + \mathcal{L}_{\text{neighbor}} \quad (6)$$

4 Experiments

4.1 Experiment Settings

Corpus We simulate a realistically low-resource setting by selecting only 2,048 parallel sentence pairs from the Tatoeba (Tiedemann, 2020) training set, which corresponds to the typical data scale of Tatoeba's lowest-resource language subset—generally consisting of only several thousand sentence pairs per language. Additionally, we collect 2 million Swahili (sw) sentences as unlabeled corpora for neighbor mining, sampled from the remaining corpus excluding the selected 2,048 pairs.

Base Encoder We use InfoXLM as the base multilingual encoder, consisting of 12 Transformer layers with a hidden size of 768.

Neighbor Mining To retrieve semantically similar neighbors in our experiments, we use sentence embeddings obtained via mean pooling over the 9th-layer hidden states of a pretrained multilingual encoder. This choice is motivated by observations in the InfoXLM (Chi et al., 2021) study, which found that representations from mid-to-late encoder layers—particularly layers 7 through 11—consistently achieved around 80% top-1 accuracy on the Tatoeba cross-lingual retrieval benchmark, indicating their effectiveness in capturing sentence-level semantics. Sentence embeddings will be ℓ_2-normalized, and then cosine similarity is used to identify the top-k nearest neighbors from the 2 million Swahili unlabeled sentences.

Hyperparameters We set the additive margin $m = 0.3$, contrastive loss temperature 0.05, and MoCo momentum 0.995. The weighted layer pooling (WLP) aggregates the last 4 hidden layers. The

projection head consists of two linear layers (hidden size → 512 → ReLU → 128). Neighbor counts are set as $k = 2$ and $k = 7$ for evaluation. Batch size is 32, queue size is 2,048, learning rate is 2e-5, and training proceeds for 30 epochs.

4.2 Evaluation

Cross-Lingual Sentence Retrieval We evaluate on the Tatoeba (Tiedemann, 2020) and FLORES (Goyal et al., 2022) benchmarks for multilingual sentence retrieval. Specifically, we use our model to encode sentences, and for each source sentence, retrieve the nearest sentence from the full target set. We then evaluate top-1 retrieval accuracy, based on whether the retrieved sentence is the exact translation. We conduct bidirectional evaluations (en→sw and sw→en).

Zero-Shot Cross-Lingual Transfer Tasks We further assess the zero-shot cross-lingual transfer capabilities of our model on classification and question answering tasks:

- **Cross-Lingual Classification:** We evaluate on the MasakhaNEWS dataset (Adelani et al., 2023), a multilingual news topic classification benchmark covering 16 languages. The model is trained on the English train set and tested zero-shot on the Swahili test set.
- **Cross-Lingual Question Answering:** We evaluate on KenSwQuAD (Wanjawa et al., 2023) and SD-QA (Faisal et al., 2021), which contain Swahili QA benchmarks where question answers are extracted from a given context. Following the MLQA (Lewis et al., 2020) setup, we finetune our model on 12K English QA pairs sampled from SQuAD (Rajpurkar et al., 2016) and evaluate its zero-shot performance on the three Swahili QA datasets.

4.3 Results

We compare the following models:

- **Base Encoder:** InfoXLM without additional training.
- **Vanilla Contrastive:** Contrastive pretraining without neighbor augmentation.
- **NeighXLM (k=2):** Neighbor-augmented contrastive pretraining with $k = 2$.
- **NeighXLM (k=7):** Neighbor-augmented contrastive pretraining with $k = 7$.

Cross-Lingual Representation Alignment Figures 3 and 4 illustrate that NeighXLM consistently outperforms both the vanilla contrastive model and the base encoder across nearly all layers on the Tatoeba and FLORES benchmarks. While standard contrastive learning already yields notable improvements over the base encoder, NeighXLM further enhances retrieval accuracy by incorporating neighborhood-based contrastive signals—particularly in the higher layers (e.g., L10–L12). Detailed results are in Appendix A. Remarkably, both NeighXLM variants (k=2 and k=7) demonstrate consistently strong performance, suggesting that the method maintains stable performance across different neighborhood sizes. These results in bi-directional sentence retrieval underscore NeighXLM's ability to effectively bridge the semantic gap across languages and promote more aligned cross-lingual representations.

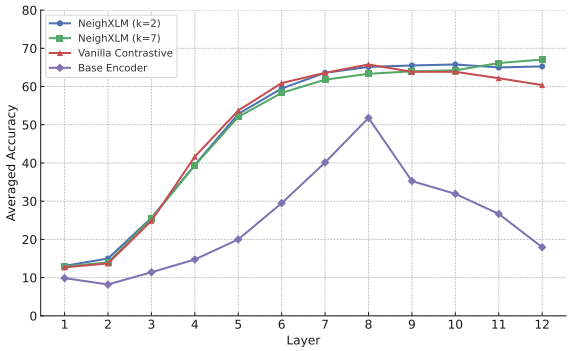

Figure 3: Layer-wise Retrieval Accuracy on Tatoeba (Averaged over en→sw and sw→en)

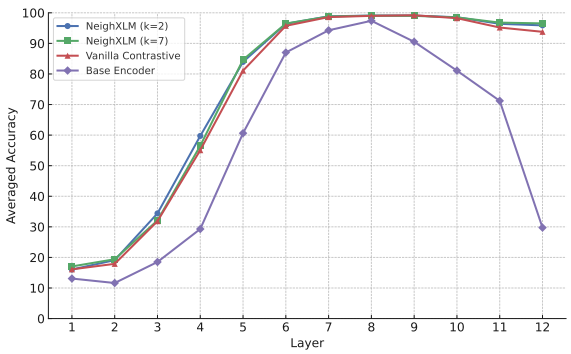

Figure 4: Layer-wise Retrieval Accuracy on FLORES (Averaged over en→sw and sw→en)

Zero Shot Cross-Lingual Classification As shown in Table 1, NeighXLM (k=2) achieves the best performance in the entertainment (F1 = 0.553) and technology (F1 = 0.711) categories.

Model	business	entertainment	health	politics	sports	technology	Avg
Base Encoder	**0.685**	0.533	**0.845**	0.804	**0.965**	0.548	0.730
Vanilla Contrastive	0.667	0.485	0.828	0.788	0.960	0.639	0.728
NeighXLM (k=7)	0.655	0.516	0.835	**0.812**	0.960	0.603	0.730
NeighXLM (k=2)	0.621	**0.553**	0.817	0.796	0.949	**0.711**	**0.741**

Table 1: F1 scores on MasakhaNEWS.

NeighXLM (k=7) performs best in the politics category (F1 = 0.812). Interestingly, the Base Encoder (InfoXLM) achieves the highest F1 scores in the business and health categories. We attribute this to topic bias in the pretraining corpora—specifically, the pretraining data used for our model differs from that of the Base Encoder, potentially leading to imbalanced topic coverage and performance variation across categories. Overall, NeighXLM (k=2) achieves the highest macro-average F1 score of 0.741, indicating its strong and consistent performance across all categories.

Zero Shot Cross-Lingual Question Answering As shown in Table 2, the NeighXLM variant with k=2 achieves the best overall performance, reaching the highest F1 and EM scores on both KenSwQuAD (49.96 / 35.76) and SD-QA (57.34 / 47.66).

Model	KenSwQuAD		SD-QA	
	F1	EM	F1	EM
Base Encoder	49.06	35.69	55.08	44.02
Vanilla Contrastive	48.27	34.37	56.39	45.47
NeighXLM (k=7)	49.28	34.75	55.72	44.02
NeighXLM (k=2)	**49.96**	**35.76**	**57.34**	**47.66**

Table 2: Results on KenSwQuAD and SD-QA.

4.4 Analysis and Discussion

Robustness to Neighbor Quality Although our neighbor search is conducted on a relatively modest pool of 2 million unlabeled sentences, both k=2 and k=7 settings lead to consistent performance gains. Manual inspection reveals that some of the more distant neighbors can be of lower semantic quality, yet the k=7 variant still performs comparably to k=2 across most tasks. This suggests that our inverse rank-based weighting mechanism plays a crucial role in mitigating the impact of noisy or less relevant neighbors, thereby enhancing the overall robustness of the model.

Importance of Neighbor Augmentation Training contrastive models with extremely limited parallel data presents significant challenges, often resulting in unstable optimization and overfitting. As evidenced by our experiments, the Vanilla Contrastive baseline—which does not incorporate neighbor augmentation—performs poorly on both the MasakhaNEWS classification and KenSwQuAD question answering tasks, in some cases even underperforming the base encoder. This underscores the limitations of contrastive objectives when applied in low-resource settings without sufficient positive supervision. By contrast, our proposed method, NeighXLM, enriches the training signal by incorporating semantic neighbors mined from unlabeled corpora as additional positive examples. This augmentation not only compensates for the lack of labeled supervision, but also mitigates overfitting and semantic space collapse by supplying more abundant and diverse positive examples, which improve coverage in the representation space.

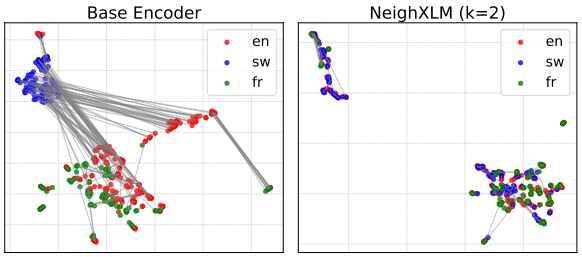

Figure 5: UMAP projection of sentence embeddings from Tatoeba (en–sw, en–fr). Each point represents a sentence, and lines connect translation pairs.

Representation Visualization We sample 100 sentence pairs each from the English–Swahili and English–French subsets of the Tatoeba benchmark. For each sentence, we compute its embedding by applying mean pooling over the final four layers of the encoder. The resulting representations are then projected to two dimensions using UMAP (McInnes et al., 2018), and visualized in Figure 5. Each point corresponds to a sentence, with lines connecting translation pairs. The visualization clearly shows that NeighXLM pro-

motes semantic clustering across languages, rather than forming clusters based on language identity—supporting its goal of enhancing cross-lingual transfer. Notably, with the base encoder, typologically similar languages like English and French already exhibit partial semantic alignment, while typologically distant languages such as Swahili are clustered strictly by language. In contrast, NeighXLM brings sentences from all three languages together based on meaning, indicating stronger and more consistent cross-lingual alignment. This language-based clustering in the base encoder also highlights a key limitation of the cross-lingual neighbor mining strategy proposed by Keung et al. (2020): selecting neighbors based on encoders that have not been aligned cross-lingually may capture superficial linguistic similarity rather than true semantics, leading to biased and less effective alignment.

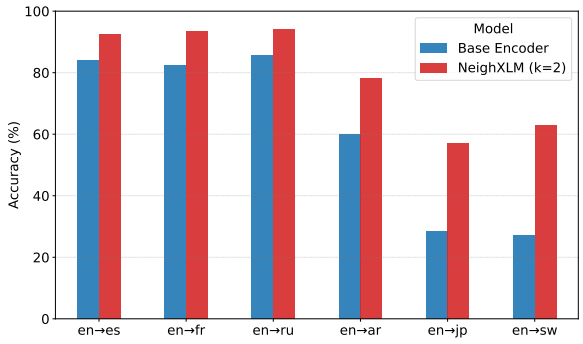

Figure 6: Sentence retrieval accuracy on additional Tatoeba language pairs (en–xx).

Preservation of Multilingual Space Beyond improving transfer to Swahili, NeighXLM does not degrade representation quality for other languages, nor does it collapse the overall multilingual semantic space. To verify this, we evaluate sentence embedding quality on several additional Tatoeba language pairs (en–xx), sampling up to 2000 sentence pairs per language. Using mean pooling over the last four layers and evaluating sentence retrieval accuracy, we observe that performance on other languages consistently improves, rather than merely remaining stable—likely because the contrastive queue loss sharpens the English representation space by pushing it away from negatives, indirectly benefiting retrieval tasks that involve English. As shown in Figure 6, this suggests that NeighXLM selectively strengthens target-language alignment while preserving or enhancing general multilingual capabilities.

Exploring Alternative Negative Queue Interactions We also experimented with alternative designs for the negative queue. Specifically, we augmented the current loss by adding an additional objective that pushes both positives and neighbor examples away from the negative queue samples. Detailed results across all evaluation tasks are provided in Appendix A; we refer to this setting as NeighXLM (allvsqueue). Overall, this variant did not lead to improved performance. Since the model already receives sufficient negative supervision through the contrastive loss, further increasing negative signals reduces the relative impact of our added positive neighbor supervision. This shift in balance diminishes the intended benefits of neighborhood-based learning, making it an inefficient modification.

5 Conclusion

In this work, we propose NeighXLM, a neighbor-augmented contrastive pretraining framework for improving cross-lingual transfer in low-resource settings. By leveraging intra-language semantic relations to mine high-quality neighbors, our method enriches supervision beyond limited parallel data and enhances cross-lingual alignment. Experiments show that NeighXLM consistently improves retrieval and zero-shot transfer performance.

6 Limitations

Although NeighXLM consistently improves performance in low-resource settings, several limitations remain.

Dependence on Unlabeled Corpora While our method removes the need for translation systems or human annotations, it still requires access to sufficient unlabeled corpora in the target language. The extent to which parallel supervision can be augmented via neighbor mining depends on the size and diversity of this corpus. For extremely low-resource languages with limited monolingual data, neighbor mining may be less effective.

Simulated Low-Resource Setting We do not use languages that are low-resource in practice in our experiments, because such languages often lack evaluation benchmarks, making it impossible to assess the performance improvements of our method. Instead, we choose Swahili, which is relatively low-resource, typologically distant from English, and

has limited but usable evaluation datasets. To simulate data scarcity, we use only a small subset of its parallel data. However, Swahili still has substantial unlabeled corpora and has been partially observed during base encoder pretraining, meaning the initial semantic space for Swahili is already of adequate quality. This gives our method a better starting point than it would have in genuinely low-resource languages that lack both labeled and unlabeled data.

References

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, and 1 others. 2023. Masakhanews: News topic classification for african languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159.

Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. 2021. Syntax-augmented multilingual bert for cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. Sd-qa: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Junxian He, Zhisong Zhang, Taylor Berg-Kirkpatrick, and Graham Neubig. 2019. Cross-lingual syntactic transfer through unsupervised adaptation of invertible projections. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3211–3223.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494.

Tao Ji, Yong Jiang, Tao Wang, Zhongqiang Huang, Fei Huang, Yuanbin Wu, and Xiaoling Wang. 2021. Word reordering for zero-shot cross-lingual structured prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4109–4120.

Tao Ji, Yuanbin Wu, and Xiaoling Wang. 2023. Typology guided multilingual position representations: Case on dependency parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13524–13541.

Phillip Keung, Julian Salazar, Yichao Lu, and Noah Smith. 2020. Unsupervised bitext mining and translation via self-trained contextual embeddings. *Transactions of the Association for Computational Linguistics*, 8:828–841.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. Mlqa: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. Veco: Variable and flexible cross-lingual pre-training for language understanding and generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3980–3994.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Jörg Tiedemann. 2020. The tatoeba translation challenge–realistic data sets for low resource and multilingual mt. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182.

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133.

Barack W Wanjawa, Lilian DA Wanzare, Florence Indede, Owen McOnyango, Lawrence Muchemi, and Edward Ombui. 2023. Kenswquad—a question answering dataset for swahili low-resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–20.

Linjuan Wu and Weiming Lu. 2023. Struct-xlm: A structure discovery multilingual language model for enhancing cross-lingual transfer through reinforcement learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3419.

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.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075.

Yinfei Yang, Gustavo Hernandez Abrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. *arXiv preprint arXiv:1902.08564*.

Zhen-Ru Zhang, Chuanqi Tan, Songfang Huang, and Fei Huang. 2023. Veco 2.0: Cross-lingual language model pre-training with multi-granularity contrastive learning. *arXiv preprint arXiv:2304.08205*.

A Results for each task

Model	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12
						sw → en						
Base Encoder	7.69	7.69	12.31	20.26	21.03	26.92	39.49	49.49	33.85	27.18	21.03	15.90
Vanilla Contrastive	10.51	12.05	22.82	42.82	52.56	59.23	62.82	64.36	63.59	62.31	61.54	60.00
NeighXLM (allvsqueue)	10.26	13.08	26.67	40.00	51.54	60.26	64.10	66.41	64.36	62.05	62.56	60.51
NeighXLM (k=7)	10.51	12.05	24.10	40.26	51.03	57.95	60.51	62.56	63.08	63.59	64.87	66.41
NeighXLM (k=2)	10.26	13.85	26.15	41.28	52.31	59.49	63.08	64.87	64.87	64.87	65.90	66.15
						en → sw						
Base Encoder	12.05	8.72	10.51	9.23	18.97	32.05	40.77	54.10	36.67	36.67	32.31	20.00
Vanilla Contrastive	14.87	15.38	26.92	40.51	54.87	62.56	64.36	67.18	64.10	65.38	62.82	60.77
NeighXLM (allvsqueue)	15.64	14.87	24.62	37.44	53.33	63.08	65.90	67.18	62.31	64.62	62.56	61.79
NeighXLM (k=7)	15.38	15.90	26.92	38.46	53.08	58.72	63.08	64.10	64.87	64.87	67.44	67.69
NeighXLM (k=2)	15.90	16.15	25.13	37.69	53.33	59.49	64.10	65.38	66.15	66.67	64.10	64.36
						bi-directional avg						
Base Encoder	9.87	8.21	11.41	14.74	20.00	29.49	40.13	51.79	35.26	31.92	26.67	17.95
Vanilla Contrastive	12.69	13.72	24.87	41.67	53.72	60.90	63.59	65.77	63.85	63.85	62.18	60.38
NeighXLM (allvsqueue)	12.95	13.97	25.64	38.72	52.44	61.67	65.00	66.79	63.33	63.33	62.56	61.15
NeighXLM (k=7)	12.95	13.97	25.51	39.36	52.05	58.33	61.79	63.33	63.97	64.23	66.15	67.05
NeighXLM (k=2)	13.08	15.00	25.64	39.49	52.82	59.49	63.59	65.13	65.51	65.77	65.00	65.26

Table 3: Layer-wise retrieval accuracy on Tatoeba.

Model	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12
						sw → en						
Base Encoder	7.32	9.23	17.45	39.42	60.68	87.46	94.08	98.19	90.57	72.72	58.78	19.96
Vanilla Contrastive	9.03	10.03	22.27	54.66	80.54	96.19	98.70	98.80	99.20	98.29	95.39	93.38
NeighXLM (allvsqueue)	7.72	9.33	20.36	53.86	83.15	97.29	98.90	99.00	98.80	97.39	94.98	91.88
NeighXLM (k=7)	8.32	10.33	22.27	56.77	84.25	97.19	99.40	99.30	99.20	98.60	96.69	95.59
NeighXLM (k=2)	8.53	10.33	23.97	59.08	84.45	97.19	99.20	99.10	99.10	98.09	96.09	95.19
						en → sw						
Base Encoder	18.86	14.04	19.56	19.16	60.58	86.56	94.48	96.59	90.47	89.47	83.65	39.52
Vanilla Contrastive	23.17	25.78	41.42	55.47	81.64	95.29	98.50	99.20	99.20	98.19	94.98	94.18
NeighXLM (allvsqueue)	24.37	26.98	41.22	56.87	82.75	96.29	98.09	99.20	98.40	98.40	95.69	93.88
NeighXLM (k=7)	25.78	28.49	41.93	56.27	85.06	95.79	98.19	98.80	98.80	98.40	96.89	97.49
NeighXLM (k=2)	23.57	27.88	44.83	60.38	83.45	95.59	98.50	99.10	99.10	99.00	96.69	96.59
						bi-directional avg						
Base Encoder	13.09	11.63	18.51	29.29	60.63	87.01	94.28	97.39	90.52	81.09	71.21	29.74
Vanilla Contrastive	16.10	17.90	31.85	55.07	81.09	95.74	98.60	99.00	99.20	98.24	95.19	93.78
NeighXLM (allvsqueue)	16.05	18.15	30.79	55.37	82.95	96.79	98.50	99.10	98.60	97.89	95.34	92.88
NeighXLM (k=7)	17.05	19.41	32.10	56.52	84.65	96.49	98.80	99.05	99.00	98.50	96.79	96.54
NeighXLM (k=2)	16.05	19.11	34.40	59.73	83.95	96.39	98.85	99.10	99.10	98.55	96.39	95.89

Table 4: Layer-wise retrieval accuracy on FLORES.

Model	KenSwQuAD		SD-QA	
	F1	**EM**	**F1**	**EM**
Base Encoder	49.06	35.69	55.08	44.02
Vanilla Contrastive	48.27	34.37	56.39	45.47
NeighXLM (allvsqueue)	49.14	35.23	57.03	46.31
NeighXLM (k=7)	49.28	34.75	55.72	44.02
NeighXLM (k=2)	49.96	35.76	57.34	47.66

Table 5: Results on KenSwQuAD and SD-QA.

Model	business	entertainment	health	politics	sports	technology	Avg
Base Encoder	0.685	0.533	0.845	0.804	0.965	0.548	0.730
Vanilla Contrastive	0.667	0.485	0.828	0.788	0.960	0.639	0.728
NeighXLM (allvsqueue)	0.678	0.556	0.831	0.792	0.959	0.583	0.733
NeighXLM (k=7)	0.655	0.516	0.835	0.812	0.960	0.603	0.730
NeighXLM (k=2)	0.621	0.553	0.817	0.796	0.949	0.711	0.741

Table 6: F1 scores on MasakhaNEWS.