

Multilingual Representation Distillation with Contrastive Learning

Weiting Tan^{♣*} Kevin Heffernan[♡] Holger Schwenk[♡] Philipp Koehn^{♣♡}

[♣]Johns Hopkins University

[♡]Meta AI Research

{wtan12, phi}@jhu.edu {kevinheffernan, schwenk}@meta.com

Abstract

Multilingual sentence representations from large models encode semantic information from two or more languages and can be used for different cross-lingual information retrieval and matching tasks. In this paper, we integrate contrastive learning into multilingual representation distillation and use it for quality estimation of parallel sentences (i.e., find semantically similar sentences that can be used as translations of each other). We validate our approach with multilingual similarity search and corpus filtering tasks. Experiments across different low-resource languages show that our method greatly outperforms previous sentence encoders such as LASER, LASER3, and LaBSE.

1 Introduction

With the rise of neural networks, high-dimensional word-level sequence representations play an important role in almost any natural language processing task. Contextual representations from large pre-trained language models (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019; CONNEAU and Lample, 2019) have shown advantages compared to earlier static embeddings (Mikolov et al., 2013; Pennington et al., 2014). However, they are not pre-trained with sentence-level objectives and their representations of two different sentences cannot be easily used to indicate semantic similarity. To encode sentence-level information, LASER (Artetxe and Schwenk, 2019b) pools a sentence embedding from the encoder and feeds it to the decoder. Another approach is to use siamese-structured models where two identical encoders are used to represent sentences of similar meaning and are trained to push their representations close to each other. Sentence-Transformers (Reimers and Gurevych, 2019) are siamese-structured models that are initialized with pre-trained large models like BERT

* Work was done during an internship at Meta AI Research.

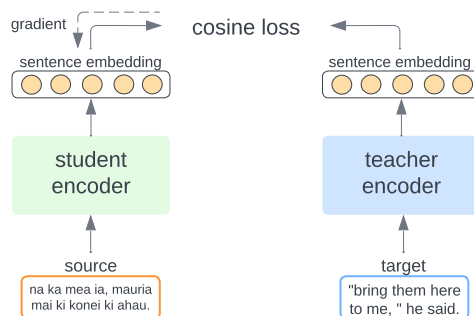


Figure 1: Student-Teacher Distillation in LASER3.

(Devlin et al., 2019) or Roberta (Liu et al., 2019). After fine-tuning, Sentence-Transformers improve their sentence representation for cross-lingual tasks. Besides fine-tuning with identical (siamese) encoders, distillation can be used to retrieve better representations. Reimers and Gurevych (2020a) extends a monolingual sentence representation into a multilingual representation with model distillation. Similarly, Heffernan et al. (2022) proposed LASER3, a student-teacher architecture that distills information from a pre-trained teacher encoder to a student encoder in new languages. As shown in Figure 1, the distillation process updates the student encoder with the gradient from the cosine loss. Note that it freezes the parameters of the teacher encoder, which is already pre-trained on the target language (English in our case). Therefore, the target sentence embedding is fixed and the corresponding source embedding is aligned with the target embedding.

In this paper, we focus on the quality estimation of parallel sentences in low-resource languages, which requires models to distinguish similar and dissimilar sentence pairs. Contrastive learning is helpful because its objective not only aligns similar sentences' representations but also pushes away representations from dissimilar sentences, which

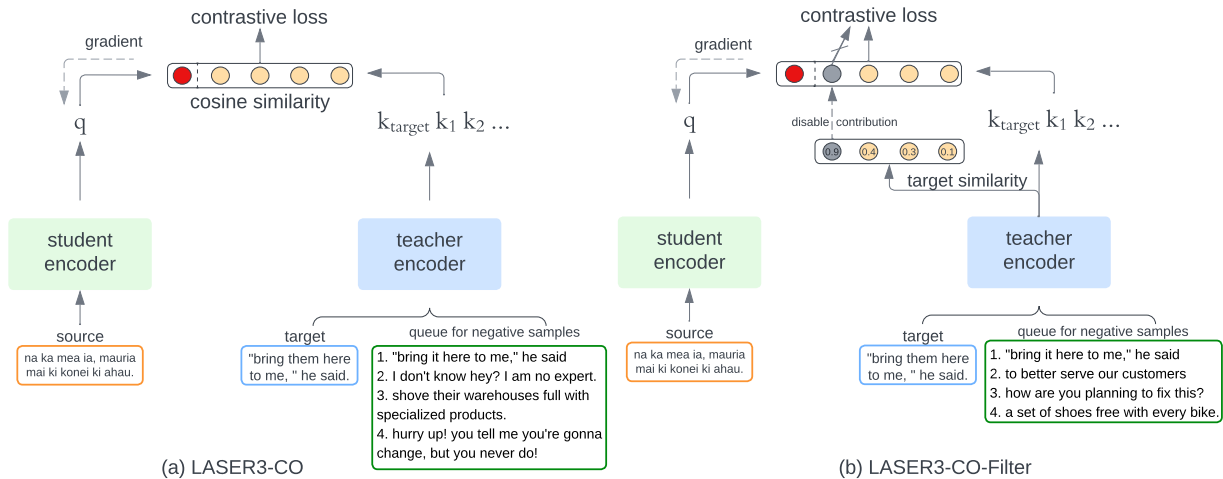


Figure 2: Visual Explanation of LASER3-CO (vanilla) and LASER3-CO-Filter (filtered). Source and target come from input bitexts dataset and queue is constructed with earlier batch’s samples. In the (b) filtered version, a pre-filtering mechanism is employed to filter out extremely negative samples from the queue.

enables the model to be more confident in similar sentences. Additionally, contrastive learning is a form of self-supervision that fits well in our low-resource setting where only a limited amount of clean data is available.

In practice, we integrate contrastive learning into the distillation-based architecture from LASER3 and use our contrastive distillation method to train encoders for low-resource languages. Inspired by He et al. (2020) and Wu et al. (2018), we used a queue to store negative samples as self-supervision to train better encoders. We also employed a pre-filtering mechanism to find hard negative samples and showed that they benefit the distillation of sentence representations.

To evaluate different encoders, we rely on multilingual similarity search and corpus filtering tasks. In the multilingual similarity search task, we encode all source and target sentences and use a cosine-based similarity metric called margin score (Artetxe and Schwenk, 2019a) to pair source and target sentences. In the corpus filtering task, a mined noisy parallel corpus is given and we use the encoder with margin score to compute a similarity score for each sentence pair. Then we filter out the low-score pairs and use the remaining corpus to train a neural machine translation model whose performance is evaluated with BLEU (Papineni et al., 2002). Compared to previous works, we observe consistent improvement when using contrastive distillation. We also compared our ap-

proach with another simple but effective data augmentation technique, back-translation, and show that contrastive distillation achieves tied or better performance with less data.

2 Approach

To train high-quality sentence representation, we integrate contrastive learning into sentence representation distillation and our contrastive distillation method is visualized in Figure 2. The motivation for using contrastive learning is two-fold:

- The self-supervision from contrastive learning helps representation learning in low-resource settings.
- Contrastive learning enables models to recognize similar and dissimilar sentences, which is crucial for filtering out noisy sentence pairs.

2.1 LASER3-CO

This architecture corresponds to Figure 2 (a). We name it LASER3-CO since it integrates contrastive learning to LASER3’s distillation pipeline. We are inspired by He et al. (2020) to use a queue to store contrastive (negative) samples. The negative samples are used to train encoders so that sentences of different meaning have dissimilar representations. During training, we use previous batches’ target language sentences as negative samples and when there are too many negative samples in the queue, we remove the samples from the earliest

batches. Our LASER3-CO approach has the following steps:

- Pre-train LASER to use as the teacher encoder θ_t on high resource languages such as English. Randomly initialize the student encoder and perform distillation with the teacher encoder. After distillation, we obtain the pre-trained student encoder θ_s .
- Fine-tune θ_s using queue(memory)-based¹ contrastive learning. For each input source sentence x and target sentence y , encode them with student and teacher encoders respectively and we have their representations $q = \theta_s(x)$, $k_{\text{target}} = k_+ = \theta_t(y)$ (we use "+" as an abbreviation for the positive target sentence). We also encode all (N) negative sentences in the queue and have their representation $k_i = \theta_t(\text{queue}_i)$, $i \in [1, N]$.
- Perform normalization on $q, k_i, i \in [0, N]$ and then compute the contrastive loss using infoNCE (van den Oord et al., 2018)

$$\mathcal{L} = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=1}^N \exp(q \cdot k_i / \tau)} \quad (1)$$

Here τ is the temperature parameter² that controls the strength of the penalty.

- Update student encoder θ_s with loss. Enqueue the most recent target language (English) sentences and dequeue the earliest sentences if the queue size exceeds the limit (N).

In LASER3-CO, our training process is very similar to MOCO (He et al., 2020), though we do not use the momentum update for the teacher. Instead, we freeze the pre-trained teacher (LASER in our case) that already has a high-quality representation of English (or any other pivot language), so that English sentences are encoded consistently during distillation. Then we try to align the representation of the student encoder to the teacher, only allowing gradients to flow through the student encoder.

2.2 LASER3-CO-Filter

This architecture corresponds to Figure 2 (b). The motivation is that previous work (Robinson et al.,

¹We use queue size N=4096 throughout experiments.

²Empirically we found $\tau = 0.05$ (widely used in previous literature) works well.

2020) has shown hard negatives improve representation. In our task, for each parallel sentence pair (x, y) , the hard negatives would be sentences y' that are hard to distinguish from the true target y (we can also find hard negatives x' in the source language, but because our teacher encoder has better representation for the target language, English, we decide to focus on target-side hard negatives only). Hard negatives are beneficial because they force the model to learn more complex features to distinguish them from the true target sentence.

To find more hard-negatives for contrastive fine-tuning, we change LASER3-CO in two ways: (1) disable shuffling for the data loader and (2) use a pre-filtering mechanism to filter out bad samples from the queue (we name this model LASER3-CO-Filter where Filter refers to the pre-filtering mechanism). Bad samples are hard negative samples y' that are too similar to y (e.g. "What is your name" versus "What's your name"). Treating these extremely similar sentences as negative samples would hurt the encoder's representation and we devise a pre-filtering method to filter them out.

Disable Shuffling We sort sentences by length³ and disable the shuffling of data loader so that consecutive batches contain sentences of similar length. As our queue is updated by enqueueing the most recent batch and dequeuing the earliest batch, disabling shuffling makes the queue store sentences of similar length. Because all samples in the queue are used as negative samples to be contrasted with the true target sentence and because they are of a similar length, it is much more likely that we find hard negatives (quantitative analysis provided in §5 and Figure 4).

Pre-filtering Mechanism After disabling shuffling, we show that more hard-negatives are found. However, we also found that there are many cases of extremely hard negatives (e.g. "What is your name" versus "What's your name", "Bring them here to me" versus "Bring it here to me"). Though hard negatives help contrastive fine-tuning to obtain better sentence representation, extremely hard negatives would confuse the model and incorrectly update the parameters. Therefore we employ a simple pre-filtering mechanism to prevent these extremely hard negatives from contributing to the contrastive loss: After we encode the target sentence

³Our implementation is based on Fairseq (Ott et al., 2019) which by default sorts the sentences by length

and sentences in the queue, we have the representations k_+, k_1, \dots, k_N as described in §2.1. Then we compute the cosine similarity between the true target and sentences in the queue and get similarity scores $\cos(k_+, k_i), i \in [1, N]$. For any negative sample i , if $\cos(k_+, k_i) \geq \sigma$ (where σ is a pre-defined threshold hyper-parameter),⁴ we do not use it in the InfoNCE loss. In other words, let $S = \{i : \cos(k_+, k_i) < \sigma\}$ be a set that contains no extremely hard negatives for the threshold σ , the InfoNCE loss would be computed as

$$\mathcal{L} = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i \in S} \exp(q \cdot k_i / \tau)} \quad (2)$$

In practice, we perform this filtering step with batches of samples, as detailed in appendix A. PyTorch-like pseudocode of the LASER3-CO-Filter model is shown below.

Algorithm 1 Pseudocode of LASER3-CO-Filter

```
# N: queue size;   bz: batch size
# S: size of set without extremely hard negatives
# t: temperature;  h: filtering threshold
# emb_dim: embedding dimension
# einsum: einsum function available in PyTorch
# queue: stores N earlier target embed
# teacher: pre-trained LASER encoder
# student: LASER3 student encoder
# filter: function that filters out extremely hard
#         negatives (details available in appendix A)

# freeze teacher encoder
teacher.params.requires_grad = False
for (source, target) in loader:
    src_emb = normalize(student(source), dim=1)
    tgt_emb = normalize(teacher(target), dim=1)
    neg_emb = normalize(teacher(queue), dim=1)
    #----- Begin Pre-filtering -----
    # bz x N (mask for indices to be kept)
    mask = (tgt_emb @ neg_emb.T < h).long()
    # bz x S x emb_dim
    neg_emb = filter(neg_emb, mask)
    #----- End Pre-filtering -----
    #positive logits: bz x 1
    l_pos = src_emb @ tgt_emb.T
    # negative logits: bz x S
    l_neg = einsum("bh,bsh->bs", src_emb, neg_emb)
    logits = cat([l_pos, l_neg], dim=1)
    loss = CrossEntropyLoss(logits/t, zeros(bz))
    loss.backward()
    update(student)
    enqueue(queue, tgt_emb)
    if queue.size > N:
        dequeue(queue)
```

3 Experiment on Low-resource Languages

3.1 Dataset

For training, we use two sources of data: existing clean data and back-translation data.

⁴We use $\sigma = 0.9$ throughout our experiments. Effects of different σ values are analyzed in Figure 4 (right)

ISO	Language	CB[k]	BT[k]
khm	Khmer	536	2451
pus	Pashto	48	3000
npi	Nepali	533	3000
sin	Sinhala	752	3000

Table 1: The number of (thousands of) sentences available for different languages in clean bitext (CB column) and back-translation (BT column) datasets. We cap back-translation data to 3 million lines maximum.

Clean data Our clean data comes from publicly available datasets⁵ such as *jw300*, *bible* (Christodoulopoulos and Steedman, 2015), *tatoeba* (Tiedemann, 2012), *wikimedia* (Tiedemann, 2012), *gv* (Tiedemann, 2012), *tico19* (Tiedemann, 2012), *ted20* (Reimers and Gurevych, 2020b), *qed* (Abdelali et al., 2014; Tiedemann, 2012), and *os* (Tiedemann, 2012).⁶ We group these data sources together and call them clean data.

Back-translation data Back-translation is a simple yet effective data augmentation technique that could benefit the training of sentence encoders. To generate back-translation data, we need a translation model and monolingual data. We retrieve monolingual data from web data provided by CommonCrawl⁷ and ParaCrawl (Bañón et al., 2020), following the pre-processing pipeline (language ID and heuristic filtering) from NLLB (NLLB Team et al., 2022). Then we use the 1.3B-parameter NLLB200-Dense⁸ model (NLLB Team et al., 2022) to translate low resource language’s monolingual data that we collected from web. We choose NLLB200 because it achieves state-of-the-art translation performance so far for most low-resource languages. Once we generate synthetic data, we can use them as a parallel corpus to distill sentence representation. In practice, we cap the amount of back-translation data at 3 million lines.

3.2 Languages & Models

We target four low-resource languages (Khmer, Pashto, Nepali, and Sinhala) in our experiments as shown in Table 1. For each of these languages,

⁵All available online, mostly taken from OPUS: <https://opus.nlpl.eu/>

⁶<https://www.opensubtitles.org/>

⁷In wet format, <https://commoncrawl.org>

⁸Open-sourced at: <https://github.com/facebookresearch/fairseq/tree/nllb>

we have trained several models shown below:

- LASER (Artetxe and Schwenk, 2019b): we directly take the model checkpoint available publicly⁹ without any further training.
- LASER3 (Heffernan et al., 2022): we follow Heffernan et al. (2022) to distill an encoder for each language using **clean data**. The teacher encoder is LASER and the student encoder is a randomly initialized transformer’s encoder.
- LASER3-BT: same as the LASER3 model except that we distill the encoder using both **clean data** and **back-translation data**. Though back-translation is already widely used, our improvement from this model provides another instance where data augmentation from a large pre-trained multilingual translation model is effective.
- LASER3-CO: we fine-tune the LASER3 model following §2.1 with **clean data**.
- LASER3-CO-Filter: we fine-tune the LASER3 model following §2.2 with **clean data**.

3.3 Multilingual Similarity Search

Multilingual similarity search evaluates the quality of representation from trained encoders by aligning source and target sentences. We use FLORES (Guzmán et al., 2019) devtest dataset (1012 aligned sentences) for this task because it is crafted with professional human annotation. For every source sentence x_i , we compute a similarity score called xsim between x_i and all target sentences y_j . If the target sentence with the highest similarity score is not the i^{th} sentence (namely if $i \neq \text{argmax}_j \text{xsim}(x_i, y_j)$), then the model has one incorrect match. We compute the percentage of source sentences that have an incorrect match as xsim error rate. The lower the xsim error rate, the better the model is for aligning sentences of similar meaning. Our similarity score xsim is the margin-based score (Artetxe and Schwenk, 2019a), which is widely used for scoring and filtering parallel sentences. It is defined below:

$$\text{xsim}(x, y) = \text{margin}(\cos(x, y), \sum_{z \in \text{NN}_k(x)} \frac{\cos(x, z)}{2k} + \sum_{z \in \text{NN}_k(y)} \frac{\cos(y, z)}{2k}) \quad (3)$$

⁹<https://github.com/facebookresearch/LASER>

Model	khm	pus	npi	sin	avg
LASER	14.6	67.8	27.0	0.1	27.4
LASER3	0.4	0.3	5.9	0.3	1.7
LASER3-BT	0.1	0.0	0.4	0.1	0.2
LASER3-CO	0.3	0.1	0.7	0.0	0.3
LASER3-CO-Filter	0.1	0.1	0.8	0.1	0.3

Table 2: xsim error rate (%) for Khmer, Pashto, Nepali, and Sinhala. Bolded values are the lowest error rate obtained for each language.

where x, y are source and target sentences’ representations from the model, $\text{NN}_k(x)$ denotes k nearest neighbors of x in the other language, and there are three margin functions¹⁰: $\text{absolute}(a, b) = a$, $\text{distance}(a, b) = a - b$, and $\text{ratio}(a, b) = \frac{a}{b}$.

We show our result in Table 2 and observe that LASER3 achieves a much better result than LASER (1.7% versus 27.4%), consistent with Heffernan et al. (2022)’s finding that language-specific distillation is more helpful than a language-agnostic encoder. We also show that using data augmentation (either with back-translation or with contrastive learning) further improves the performance to a near-zero error rate. In Figure 3, we provide the t-SNE visualization of embedding spaces trained by different approaches for Sinhala, and we observe that using back-translation or contrastive learning brings aligned sentences closer.

Note that in both of our contrastive-based models (LASER3-CO and LASER3-CO-Filter), we do not include back-translation data. Thus, contrastive distillation achieves tied performance compared to the model helped by a large amount of back-translation data, showing the power of self-supervision in representation learning. Additionally, back-translation and contrastive learning are orthogonal so we could train LASER3-CO/LASER3-CO-Filter with back-translation data. However, in our preliminary study, we do not see improvement from such an approach (more discussion in Appendix C) and we use only clean data for contrastive distillation throughout our experiments.

3.4 Corpus Filtering

We pick Khmer, Pashto, Nepali, and Sinhala as our targeted low-resource languages partly because they are used by WMT Corpus Filtering Shared task (Koehn et al., 2019, 2020). The shared task re-

¹⁰We follow Heffernan et al. (2022) and NLLB Team et al. (2022) to use ratio as the margin.

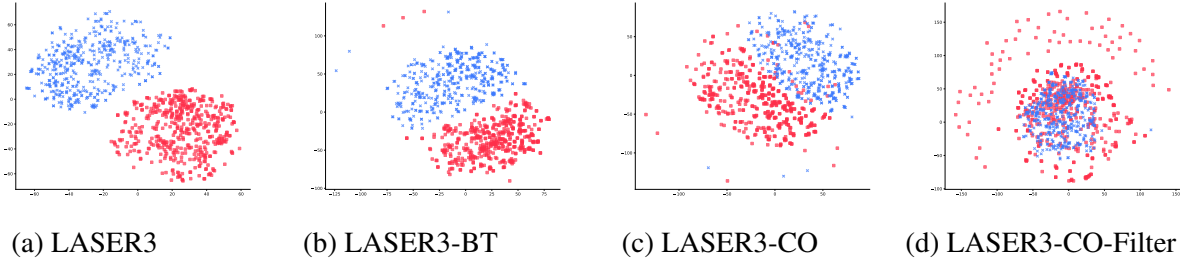


Figure 3: t-SNE visualizations of embedding space for Sinhala-English. Red points are from Sinhala sentences and blue points are from English sentences.

Model	khm	pus	npi	sin
†LASER	5	5	N/A	N/A
†WMT Best System	5	5	1	1
LASER3	7	5	1	2
LASER3-BT	7	5	2	5
LASER3-CO	7	5	2	5
LASER3-CO-Filter	7	7	2	5

Table 3: Optimal subsample size (in millions of tokens) for corpus filtering task. For most cases, 5 or 7 million tokens give the best performance except for Nepali, where only 1 or 2 million tokens seem to be useful. †: number taken from findings of WMT corpus filtering shared task (Koehn et al., 2019, 2020).

lease mined corpora (from the Paracrawl¹¹ project) for Khmer, Pashto, Nepali, and Sinhala. For each language, we first filter the noisy corpus with the language id model from fasttext¹² (Joulin et al., 2017). We filter out the aligned sentence pairs whose source sentence is predicted to be English (because we only expect the target sentence to be English). Then we filter the corpus with our sentence encoders where we rely on xsim score to rank sentences and take the top k sentences to train NMT models. In practice, we follow the shared task to use data subset of 1/2/3/5/7 million English tokens, and the optimal subset size is shown in Table 3. The highest BLEU score achieved by different systems is shown in Table 4.

From the results, we observe that LASER3-based systems have similar BLEU scores for Khmer. One possible explanation is that LASER3 distillation already gives good representation so data augmentation does not make a noticeable improvement. On average, we find that data-augmented systems (LASER3-BT, LASER3-CO, LASER3-CO-Filter) greatly improve the BLEU

Model	khm	pus	npi	sin	avg
†LASER	7.1	9.7	N/A	N/A	
†WMT Best System *	8.9	10.9	6.9	6.5	8.30
LASER3	10.1	11.7	5.1	5.8	8.18
LASER3-BT	10.1	11.8	6.2	7.7	8.95
LASER3-CO	9.9	12.3	6	7.3	8.88
LASER3-CO-Filter	10.1	12.2	6.2	7.7	9.05

Table 4: BLEU score for Khmer, Pashto, Nepali, and Sinhala. Bolded values are the highest BLEU scores obtained for each language. *: The highest number achieved by WMT19 & 20 corpus filtering task participants. †: results taken from findings of WMT19 & 20 (Koehn et al., 2019, 2020).

score compared to previous works like LASER, LASER3, and best systems in the WMT Corpus Filtering shared task. Note that some systems from the shared task make use of large pre-trained models and even an ensemble of large models, so it is a strong baseline and we are able to outperform it by an average of 0.75 BLEU score, achieving the state-of-the-art corpus filtering performance using contrastive distillation on a small amount of clean data. The corpus filtering result is also consistent with the multilingual similarity search task, where encoders with lower xsim error rate perform better.

4 Experiment on Extremely Low-resource Languages

4.1 Languages & Dataset

In §3, we experiment on low-resource languages **individually** and show the improvement from contrastive distillation. There are many languages that have even less available data and are under-represented in the current literature. In our work, we focus on four extremely low-resource African languages that are extremely low-resource: Chokwe, Kamba, Kimbundu, and Umbundu. The amount of available clean and back-translation data is shown in Table 5 (with the same collection pro-

¹¹<https://paracrawl.eu/>

¹²<https://fasttext.cc/docs/en/language-identification.html>

ISO	Language	CB[k]	BT[k]
cjk	Chokwe	40	166
kam	Kamba	58	74
kmb	Kimbundu	101	140
umb	Umbundu	234	366

Table 5: the number of (thousands of) sentences available for different African languages in clean bitext (CB column) and back-translation (BT column) datasets. We cap back-translation data to 3 million lines maximum.

cess described in §3.1). These four languages are from the Benue-Congo language family which has 25 languages in total. The details of their names and available data size are shown in Table 8 (in the appendix). Under such extremely low-resource settings, we show that contrastive distillation still produces state-of-the-art encoders for the multilingual similarity search task. We believe our encoders can be easily adapted to other information extraction tasks such as bitext mining through the pipeline described in CCMatrix (Schwenk et al., 2021), and we leave this to future work.

4.2 Multilingual Similarity Search

We evaluate the performances of difference encoders on the multilingual similarity search task following the same procedure in §3.3 and show `xsim` error rate in Table 6. For the language-agnostic model baseline, we replace LASER with LaBSE (Feng et al., 2022) because its pre-training data contains African languages (though not directly on the four languages we target) and achieves better performance than LASER. However, this language-agnostic model is still easily outperformed by LASER3-based models since the distillation of language-specific data is helpful.

For LASER3 and LASER3-BT, we leverage all data from the Benue-Congo family to distill the encoder. We are able to reproduce results from LASER3 (Heffernan et al., 2022) and our back-translation baseline further improves the average `xsim` error rate to 6.7%. For contrastive distillation models, we fine-tune LASER3-BT **individually**¹³ for each target language following §2. Not like previous experiments with Khmer, Pashto, etc., we fine-tuned LASER3-BT instead of LASER3 as our preliminary study showed better performance with LASER3-BT. Therefore, different languages could

¹³Consistent with our experiment in §3.3, using only clean data works the best for contrastive distillation.

Model	cjk	kam	kmb	umb	avg
LaBSE*	34.4	27.4	35.0	37.0	33.5
LASER3*	16.4	15.3	7.5	15.6	13.7
LASER3-BT	14.6	3.5	4.8	7.0	6.7
LASER3-BT-CO	9.1	1.9	1.3	4.6	4.2
LASER3-BT-CO-Filter	11.3	2.2	1.2	7.1	5.5

Table 6: `xsim` error rate (%) for Chokwe, Kamba, Kimbundu, and Umbundu. Bolded values are the lowest error rate obtained for each language. *: result taken from Heffernan et al. (2022).

have different optimal settings for data augmentation (more discussion in Appendix C), depending on the amount and quality of available data. Nevertheless, we show that our contrastive distillation framework greatly improves upon previous encoders, achieving state-of-the-art `xsim` error rates for all four languages we targeted.

5 Analysis

In this section, we focus on contrastive distillation methods described in §2 and provide an analysis of the effects of different hyper-parameters.

5.1 Effect of queue size

For all the previous experiments, we use batch size 32 and queue size 4096 to be consistent. In our preliminary work, we tried $N=512/1024/2048/4096$ for Sinhala and we do not observe a noticeable advantage for using a larger queue size. Still, we use the largest queue size that we could afford,¹⁴ 4096, throughout experiments since a larger queue size won’t hurt performance. To show that memory-based contrastive learning is effective, we compare it with in-batch fine-tuning and show the result in Table 7. We see that memory-based contrastive learning gives us an average of 0.4 BLEU improvement across four languages.

Model	khm	pus	npi	sin	avg
LASER3	10.1	11.7	5.1	5.8	8.18
LASER3-CO	9.9	12.3	6	7.3	8.88
LASER3-InBatch	9.8	11.4	5.8	6.9	8.48

Table 7: BLEU score for Khmer, Pashto, Nepali, and Sinhala. LASER3-InBatch and LASER3-CO have the same setup except that LASER3-CO uses a queue of 4096 samples to compute contrastive loss and LASER3-InBatch only uses samples from every batch to compute contrastive loss.

¹⁴We trained our models with NVIDIA A100 GPU.

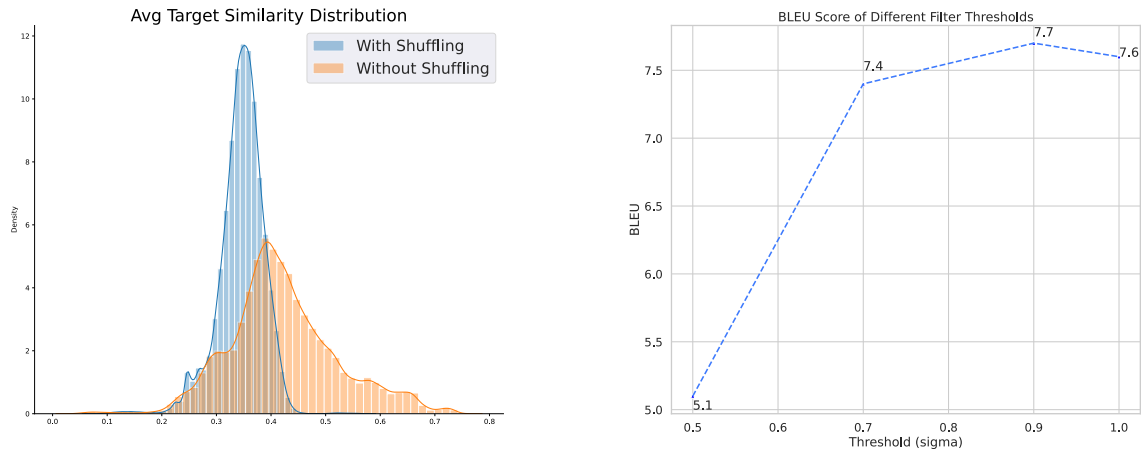


Figure 4: (left) Distribution plot of the average target similarity (of Sinhala’s clean data) with/without shuffling. The x-axis is the value of average target similarity and the y-axis is the density of a particular value. (right) LASER3-CO-Filter’s performance for Sinhala w.r.t varying pre-filter threshold σ . A low threshold filters out all hard negatives and harms training. A high threshold keeps extremely hard negatives which also undermines training.

5.2 Effect of shuffling

In our LASER3-CO-Filter approach, we disable shuffling so that the queue contains sentences of similar length and we argue that it is easier to find hard negatives this way. To quantify this fact, we compute the average target similarity for all sentences and plot the distribution in Figure 4 (left). The average target similarity for a target sentence x is

$$\text{sim}(x, \text{queue}) = \frac{1}{N} \cos(\theta_t(x), \theta_t(\text{queue}_i))$$

where θ_t is the teacher encoder, N is queue size, and queue_i is the i^{th} sample in the queue. As shown in Figure 4 (left), without shuffling, more than half samples have > 0.4 average target similarity. A higher average target similarity indicates that target sentences are more semantically related to the samples in the queue and therefore form harder negatives.

5.3 Effect of pre-filter threshold

For the LASER3-CO-Filter model, we employ a pre-filtering mechanism that prevents extremely hard examples from contributing to the loss. For our experiments, we use filter threshold $\sigma = 0.9$. In this section, we vary σ and evaluate the performance for Sinhala (we did not repeat this experiment for other languages but expect similar results). The result is shown in Figure 4 (right), which is a line plot based on our result for $\sigma = 0.5/0.7/0.9/1$. We can see that $\sigma = 0.9$ is the peak and the result

corresponds to our intuition. When σ is too high (e.g. $\sigma = 1$), extremely hard negatives (or even repetitions) are not filtered out and their contribution to contrastive loss makes training unstable, resulting in poor performance for mining. When σ is too low (e.g. $\sigma = 0.5$), too many samples in the queue are filtered out and only easy negatives (sentences drastically different from the target sentence) are left in the queue. In this case, the model does not learn meaningful representations from negative samples, resulting in poor performance as well.

6 Related Work

Multilingual Sentence Representation Various multilingual sentence representations have been proposed including multilingual BERT (Devlin et al., 2019), XLM (CONNEAU and Lample, 2019), XLM-R (Conneau et al., 2020). These models are trained with a large amount of data across dozens or hundreds of languages. They show great improvement when fine-tuned on downstream tasks. However, they are trained only with token-level objectives like masked language modeling (MLM) or translation language modeling (TLM) and their sentence representations cannot be directly used for semantic similarity tasks. To address this problem, different methods make use of pooled sentence embedding to train better encoders (Artetxe and Schwenk, 2019b; Reimers and Gurevych, 2019; Feng et al., 2022; Heffernan et al., 2022).

Back-translation Back-translation is widely used in machine translation as a data augmentation technique. [Sennrich et al. \(2016\)](#) shows that synthetic data created by translating monolingual data greatly improve English-German and Turkish-English translation models. [Edunov et al. \(2018\)](#) further investigate different decoding strategies and show that sampling works better than greedy decode or beam search for back-translation. Since back-translation’s quality depends on the translation model, [Hoang et al. \(2018\)](#) proposes iterative back-translation where translation models are further improved by the synthetic data they produce.

Contrastive Learning Contrastive learning is a form of self-supervised learning that encourages similar inputs to have close representations and dissimilar inputs to have different representations. [Chen et al. \(2020\)](#) proposes SimCLR that uses contrastive learning to improve visual representation. [Khosla et al. \(2020\)](#) integrates supervised learning into contrastive learning and [Tian et al. \(2020\)](#) uses contrastive distillation for image representation learning. Since finding high-quality contrastive samples is essential for contrastive learning, [Wu et al. \(2018\)](#) uses a memory bank to store contrastive samples while [He et al. \(2020\)](#) uses a queue with momentum update to encode contrastive samples on the fly. Though previously-mentioned works are evaluated on computer vision tasks, there is growing interest in using contrastive learning for natural language processing and various approaches have been proposed to train better sentence representation ([Giorgi et al., 2021](#); [Kim et al., 2021](#); [Peng et al., 2020](#); [Gao et al., 2021](#); [Tan and Koehn, 2022](#)).

7 Conclusion

In this paper, we integrate contrastive learning into multilingual sentence representation distillation. Through self-supervision from negatives, especially hard negatives, we obtain better encoders. Compared to other effective data augmentation techniques like back-translation, our contrastive distillation method achieves tied or better performance with less computation and data. Our experiments on multilingual similarity search and corpus filtering tasks show consistent improvement from contrastive learning and achieve state-of-the-art performance for various low-resource languages.

Limitations

This research focuses on improving cross-lingual encoders to perform quality estimation of parallel data for low-resource machine translation. However, cross-lingual representation could be helpful in many other tasks as summarized in the XTREME benchmark ([Hu et al., 2020](#); [Ruder et al., 2021](#)) which we have not experimented with. Another limitation is that we disabled shuffling to store hard negative samples in the queue. However, without shuffling, it might affect the stability of the training process (although we do not observe it in our experiments). Therefore, in future work, we will continue exploring other cost-efficient approaches to retrieve hard negatives.

Acknowledgements

We thank anonymous reviewers for their valuable comments. We also thank Alex Guo, Huda Khayrallah, Kaushik Ram Sadagopan, Simeng Sun, Guillaume Wenzek, and Haoran Xu for their helpful suggestions.

References

- Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. [The AMARA corpus: Building parallel language resources for the educational domain](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Mikel Artetxe and Holger Schwenk. 2019a. [Margin-based parallel corpus mining with multilingual sentence embeddings](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.
- Mikel Artetxe and Holger Schwenk. 2019b. [Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond](#). *Transactions of the Association for Computational Linguistics*, 7:597–610.
- Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. [ParaCrawl: Web-scale acquisition of parallel corpora](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

- 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*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.
- Christos Christodoulopoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49:375–395.
- 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.
- Alexis CONNEAU and Guillaume Lample. 2019. [Cross-lingual language model pretraining](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- 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, Minneapolis, Minnesota. Association for Computational Linguistics.
- Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. [Understanding back-translation at scale](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.
- 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, Dublin, Ireland. Association for Computational Linguistics.
- 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. [DeCLUTR: Deep contrastive learning for unsupervised textual representations](#). 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 879–895, Online. Association for Computational Linguistics.
- Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc’Aurelio Ranzato. 2019. [The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English](#). 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 6098–6111, Hong Kong, China. Association for Computational Linguistics.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. [Momentum contrast for unsupervised visual representation learning](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.
- Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. [Bitext mining using distilled sentence representations for low-resource languages](#).
- Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. [Iterative back-translation for neural machine translation](#). In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. [XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of tricks for efficient text classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. [Supervised contrastive learning](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.
- Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. [Self-guided contrastive learning for BERT sentence representations](#). 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 879–895, Online. Association for Computational Linguistics.

- Papers*), pages 2528–2540, Online. Association for Computational Linguistics.
- Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. [Findings of the WMT 2020 shared task on parallel corpus filtering and alignment](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742, Online. Association for Computational Linguistics.
- Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Pino. 2019. [Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 54–72, Florence, Italy. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). Cite arxiv:1907.11692.
- Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. [No language left behind: Scaling human-centered machine translation](#).
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. [fairseq: A fast, extensible toolkit for sequence modeling](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Wei Peng, Jianfeng Liu, Minghan Wang, Liangyou Li, Xupeng Meng, Hao Yang, and Qun Liu. 2020. [Huawei’s submissions to the WMT20 biomedical translation task](#). In *Proceedings of the Fifth Conference on Machine Translation*, pages 857–861, Online. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. [GloVe: Global vectors for word representation](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2020a. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2020b. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.
- Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. [Contrastive learning with hard negative samples](#).
- Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. [XTREME-R: Towards more challenging and nuanced multilingual evaluation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. [CCMatrix: Mining billions of high-quality parallel sentences on the web](#). 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 6490–6500, Online. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Weiting Tan and Philipp Koehn. 2022. [Bitext mining for low-resource languages via contrastive learning](#).

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive representation distillation. In *International Conference on Learning Representations*.

Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. [Representation learning with contrastive predictive coding](#). *CoRR*, abs/1807.03748.

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.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

A PyTorch-Style Pseudocode

In Algorithm 1, we provide the pythonic style pseudocode for our LASER3-CO-Filter model. Note that the pre-filtering mechanism needs special treatment when training with a minibatch. we compute the set of negative samples

$$S^j = \{i : \cos(k_+^j, k_i) < \sigma\}$$

for each target sentence j in our batch where k_+^j is the teacher encoding of j^{th} positive sample. Then we find the minimum number of negatives in the batch $M = \min_j |S^j|$. For any target sentence j that has $|S^j| > M$, we randomly remove $|S^j| - M$ samples so that it ends up with M samples. In this way, we ensure that every target sentence has the same number of legitimate negative samples after the pre-filtering mechanism and we can batch them to compute the InfoNCE loss. In the pseudocode, we abstract the implementation of pre-filtering using `filter` function.

B Corpus Filtering Threshold

In Table 3, we show the optimal size of filtered corpus to train NMT models for four languages. For most cases, 5 or 7 million tokens give the best performance (except for Nepali, where only 1 or 2 million tokens seem to be useful).

C Combine Back-translation and Contrastive Learning?

Back-translation and contrastive learning are two helpful ways for representation learning in low-resource settings. These two methods can be applied together (as we showed in our experiments on African languages) because back-translation prepares data before training while our contrastive distillation method is a fine-tuning strategy. Through our preliminary study, we made the following observations:

- Contrastive fine-tuning works best on a single language’s clean data. This is true for both low-resource languages (Khmer, Pashto, Nepali, and Sinhala) and extremely low-resource languages (Chokwe, Kamba, Kimbundu, and Umbundu). We find that using back-translation data for contrastive objectives does not make any noticeable improvement and this is probably because synthetic data generated by the pre-trained NMT model

does not produce too many hard negatives. In the case of the African languages, we tried distillation with the whole family’s clean data and it is also less effective than distillation on a single language. This is probably because the data from multiple languages interfere with each other during contrastive distillation.

- For low-resource languages like Sinhala, Pashto, etc., we find no noticeable difference between contrastive distillation with LASER3 and LASER3-BT. On the contrary, for extremely low-resource languages, we find LASER3-BT-CO performs better than LASER3-CO, and this is why we used LASER3-BT-CO in §4.2 but LASER3-CO in §3.3. We believe this is due to the quality and the amount of available clean data. In the extremely low-resource case, even the limited amount of clean data contains noise and we see LASER3-BT greatly outperform LASER3. Therefore, the distillation with pre-trained LASER3-BT result in a better model (LASER3-BT-CO or LASER3-BT-CO-Filter). However, in some low-resource settings (like our experiments for Sinhala, Pashto, Khmer, and Nepali), clean data itself could allow the contrastive method to learn good representations, therefore avoiding the need for back-translation data.

With our observation above, how should we select methods when new languages come in? We believe it mainly depends on two factors: 1) How large and how clean is the existing clean corpus 2) How is the performance of the existing pre-trained NMT model?

The first factor affects the effectiveness of contrastive learning and the second factor affects the quality of back-translation data. In fact, in our preliminary study, we do not find back-translation from a statistical machine translation system useful for African languages, because it does generate good enough synthetic data. With the advance of multilingual machine translation (MMT) models, we are able to leverage a large pre-trained model and show its effectiveness. Thus, whether to add back-translation or not mainly depends on the translation model’s quality and computation budget.

ISO	lang	CB[k]	BT[k]	ISO	lang	CB[k]	BT[k]
khm	Khmer	536	2451	pus	Pashto	48	3000
npi	Nepali	533	3000	sin	Sinhala	752	3000
ckj	Chokwe	40	166	kam	Kamba	58	74
kmb	Kimbundu	101	140	umb	Umbundu	234	366
bem	Bemba	700	828	ibo	Igbo	912	3000
kik	Kikuyu	119	95	kin	Kinyarwanda	517	3000
kon	Kongo	227	197	lin	Lingala	981	848
lua	Luba-Kasai	321	492	lug	Luganda	340	1654
nso	Northern Sotho	625	902	nya	Chewa;Nyanja	867	3000
run	Rundi	664	1910	sna	Shona	789	3000
sot	Sotho	1510	3000	ssw	Swati	116	432
swh	Swahili	2778	0	tsn	Tswana	1748	1680
tso	Tsonga	852	1423	tum	Tumbuka	264	675
xho	Xhosa	1739	3000	yor	Yoruba	517	3000
zul	Zulu	2383	3000				

Table 8: number of (thousands of) sentences available for different languages in clean bitext (CB column) and back-translation (BT column) dataset.

ISO	bible	gv	jw300	qed	tatoeba	ted20	tico19	wikimedia	os
khm	✓	✓	✓	✓	✓	✓	✓	✓	
pus				✓	✓	✓	✓	✓	
npi	✓	✓	✓	✓	✓	✓		✓	
sin	✓		✓	✓	✓	✓		✓	✓
ckj			✓						
kam			✓		✓				
kmb			✓						
umb			✓	✓	✓				

Table 9: Source of clean bitext data for different languages.