Improving Multi-lingual Alignment Through Soft Contrastive Learning

Minsu Park[*]
Yonsei University
0601p@yonsei.ac.kr

Seyeon Choi[*]
Yonsei University
seyeon717@yonsei.ac.kr

Chanyeol Choi
Linq
jacob.choi@getlinq.com

Jun-Seong Kim[†]
Linq
junseong.kim@getlinq.com

Jy-yong Sohn[†]
Yonsei University
jysohn1108@gmail.com

Abstract

Making decent multi-lingual sentence representations is critical to achieve high performances in cross-lingual downstream tasks. In this work, we propose a novel method to align multi-lingual embeddings based on the similarity of sentences measured by a pre-trained mono-lingual embedding model. Given translation sentence pairs, we train a multi-lingual model in a way that the similarity between cross-lingual embeddings follows the similarity of sentences measured at the mono-lingual teacher model. Our method can be considered as contrastive learning with *soft* labels defined as the similarity between sentences. Our experimental results on five languages show that our contrastive loss with soft labels far outperforms conventional contrastive loss with hard labels in various benchmarks for bitext mining tasks and STS tasks. In addition, our method outperforms existing multi-lingual embeddings including LaBSE, for Tatoeba dataset. The code is available at https://github.com/YAI12xLinq-B/IMASCL

1 Introduction

Learning good representations (or embeddings) of sentences and passages is crucial for developing decent models adaptive to various downstream tasks in natural language processing. Compared with the high quality mono-lingual sentence embeddings developed in recent years (Wang et al., 2023; Song et al., 2020), multi-lingual sentence embeddings have a room for improvement, mostly due to the difficulty of gathering translation pair data compared to mono-lingual data. This motivated recent trials on improving the performance of multi-lingual embeddings.

One of the prominent approaches trains the multi-lingual embeddings using contrastive learning (Zhang et al., 2022; Gao et al., 2021). Given a translation pair for different languages, this approach trains the model in a way that the embeddings for translation pairs are brought closer together, while embeddings for non-translation pairs are pushed further apart (Feng et al., 2020). Despite several benefits of this contrastive learning approach, Ham and Kim (2021) pointed out that current training method ruins the mono-lingual embedding space. To be specific, this issue arises from the fact that existing contrastive loss treats sentences that are not exact translation pairs identically (as negative pairs), irrespective of the semantic similarity of those sentences.

Another prominent approach is distilling mono-lingual teacher embedding space to a multi-lingual student model. The basic idea is, letting the multi-lingual embeddings of student models follow the mono-lingual embeddings of teacher model. This approach is motivated by the assumption that English embeddings are well constructed enough to guide immature multi-lingual embeddings. For example, Reimers and Gurevych (2020) proposed a distillation method using mean-squared-error (MSE) loss, which is shown to be effective in learning embeddings for low-resource languages. Also, Heffernan et al. (2022) used a distillation method where the teacher is the English embedding of a multi-lingual model. Unfortunately, existing distillation methods cannot fully utilize the translation pairs. Since conventional methods choose the most reliable English embeddings as the teacher model, translation pairs from non-English language parallel corpus are not fully leveraged.

In this paper, we propose a novel distillation method for improving multi-lingual embeddings, by using *soft* contrastive learning. See Figure 1. Given N translation pairs $\{(s_i, t_i)\}_{i=1}^{N}$, our method first computes the mono-lingual sentence similarity matrix from the teacher model. Each element of this similarity matrix is a continuous value. We distill such *soft* label to the cross-lingual sentence

[*]Equal contribution
[†]Corresponding authors

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 138–145

similarity matrix computed for the multi-lingual student model. In other words, the anchor similarity matrix computed from the teacher model is used as a pseudo-label for contrastive loss. Our main contributions are as follows:

- We propose a novel method of fine-tuning multi-lingual embeddings by distilling the sentence similarities measured by mono-lingual teacher models. Compared with the conventional contrastive learning which uses hard labels (either positive or negative for sentence pairs), our method chooses *soft* labels for measuring the sentence similarities.
- Compared with conventional contrastive learning and monolingual distillation method using MSE, our soft contrastive learning has much improved performance in bitext mining tasks, Tatoeba, BUCC and FLORES-200, for five different languages.
- For Tatoeba, our method outperforms existing baselines including LaBSE, LASER2 and MPNet-multi-lingual.

2 Related Works

Constructing multi-lingual embedding has been actively studied for recent years (Heffernan et al., 2022; Artetxe and Schwenk, 2019; Duquenne et al., 2023). For example, LaBSE (Feng et al., 2020) shows remarkable bitext retrieval performances, which is first pretrained with masked language modeling (MLM) (Devlin et al., 2018) and translation language modeling (TLM) (Conneau and Lample, 2019) tasks, and then fine-tuned with translation pairs using contrastive loss, i.e. translation ranking task. Also, mUSE (Yang et al., 2019) uses translation based bridge tasks from Chidambaram et al. (2018) to make a multilingual embedding space. In short, mUSE (Yang et al., 2019) is trained for translation ranking task with hard negatives.

Reimers and Gurevych (2020) introduced distilling the mono-lingual embedding of a teacher model (using sBERT (Reimers and Gurevych, 2019)) to the multi-lingual embedding of a student model (using XLM-R (Conneau et al., 2019)) with MSE loss, which enables a good bitext retrieval performance only with small amounts of parallel data. Several follow-up papers (Duquenne et al., 2023; Heffernan et al., 2022) achieved good performances by using this distillation approach. For example, Heffernan et al. (2022) successfully improved the performance on low-resource languages with the aid of the distillation approach. Compared with existing distillation methods, our work distills the similarities between sentences measured by mono-lingual embeddings, instead of directly distilling the mono-lingual embedding space of the teacher model.

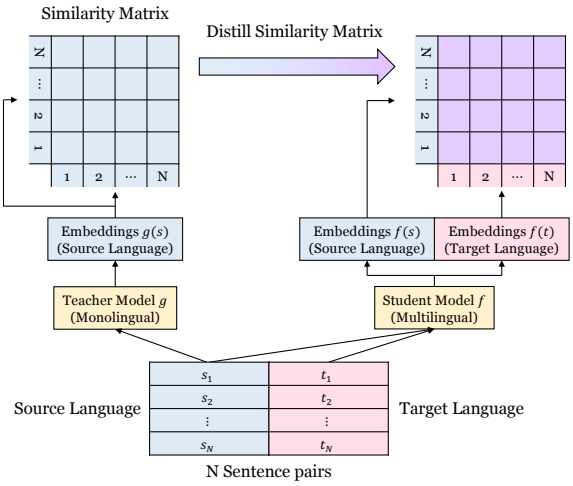

Figure 1: Overall framework of our method. Given N sentence pairs from source/target languages, we train a multi-lingual student model f by using the similarity between sentences measured by a mono-lingual teacher model g. Our contrastive loss function in Eq. 2 uses soft-label $w(i, j)$ defined in Eq. 4 and 5.

3 Proposed Method

Suppose we are given N translation pairs, denoted by $(s_1, t_1), (s_2, t_2), \cdots, (s_N, t_N)$, where s_i is the i-th sentence in the source language and t_i is the corresponding sentence in the target language. We train a multi-lingual student model f by using the similarities between mono-lingual sentences measured by a teacher model g. Here, g can be either a mono-lingual model or using only a single language from multi-lingual models. Specifically, we first use the teacher network g to measure the similarity of sentences $\{s_i\}_{i=1}^N$ in the source language. The cosine similarity between sentences s_i and s_j measured by encoder g is denoted by

$$\texttt{sim}_g(s_i, s_j) = \frac{\texttt{cos}(g(s_i), g(s_j))}{\tau},$$

where τ is the temperature parameter. Then, we train multi-lingual encoder f in a way that

$$\texttt{sim}_g(s_i, s_j) \approx \texttt{sim}_f(s_i, t_j) \approx \texttt{sim}_f(t_i, s_j) \quad (1)$$

i.e., the similarity of i-th sentence and j-th sentence is maintained across different language pairs. Such objective is reflected in our contrastive loss

$$L_{row} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} w(i,j)\log\left(\frac{e^{\texttt{sim}_f(s_i,t_j)}}{\sum_{n=1}^{N} e^{\texttt{sim}_f(s_i,t_n)}}\right) \quad (2)$$

where $w(i,j)$ is the label using similarity between s_i and s_j. The standard contrastive loss used in LaBSE (Feng et al., 2020) and mE5 (Wang et al., 2024) has the form of Eq. 2 where

$$w(i,j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

i.e., two sentences (s_i and s_j) are considered as a positive pair only if $i = j$, and labeled as a negative pair otherwise.

Since the naïve labeling method above cannot fully capture the semantic relationship between different sentence pairs, we propose following $w(i,j)$ by applying the softmax function on the similarity matrix measured at the teacher model.

$$w(i,j) = \frac{e^{\texttt{sim}_g(s_i,s_j)}}{\sum_{n=1}^{N} e^{\texttt{sim}_g(s_i,s_n)}} \quad (4)$$

$w(i,j)$, namely Priority label, calculates label based on the similarity using the anchor language sentences. In Eq 4, we assume the source language as an anchor. Note that both the source and the target language are available as an anchor language, thus we need to choose one.

Thus, we consider a variant of $w(i,j)$, namely Average label, which mixes monolingual embedding spaces of source and target language by averaging similarity.

$$w(i,j) = \frac{e^{(\texttt{sim}_g(s_i,s_j)+\texttt{sim}_g(t_i,t_j))/2}}{\sum_{n=1}^{N} e^{(\texttt{sim}_g(s_i,s_n)+\texttt{sim}_g(t_i,t_n))/2}} \quad (5)$$

In fact, this only works when the teacher model is multi-lingual, and the student encoder f trains in a following way, which is different from the Eq. 1.

$$(\texttt{sim}_g(s_i,s_j) + \texttt{sim}_g(t_i,t_j))/2 \approx \texttt{sim}_f(s_i,t_j) \approx \texttt{sim}_f(t_i,s_j)$$

The contrastive loss discussed above is unidirectional. Following the common symmetric bidirectional contrastive loss, e.g., (Radford et al., 2021), the symmetric loss using our soft label is defined as

$$L_{cross} = L_{row} + L_{col} \quad (6)$$

where L_{row} is in Eq. 2 and

$$L_{col} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} w(i,j)\log\left(\frac{e^{\texttt{sim}_f(s_i,t_j)}}{\sum_{n=1}^{N} e^{\texttt{sim}_f(s_n,t_j)}}\right).$$

Training Monolingual Space (TMS) Note that our objective L_{cross} given in Eq. 6 is to learn only the cross-lingual similarity of the student model. In addition to that, we consider learning with additional mono-lingual loss L_{mono} in Eq. 7,the distillation loss measured by the similarity between each monolingual sentence pair. This approach of using L_{mono} on top of L_{cross} is dubbed as training monolingual space (TMS). The combined loss term is shown in Eq. 8, where the parameter λ controls the balance between the cross-lingual loss and the mono-lingual loss term. Note that using TMS is orthogonal to the choice of using the Priority label or the Average label.

$$\begin{aligned} L_{mono} = &-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} w(i,j)\log\left(\frac{e^{\texttt{sim}_f(s_i,s_j)}}{\sum_{n=1}^{N} e^{\texttt{sim}_f(s_n,s_j)}}\right) \\ &+ -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} w(i,j)\log\left(\frac{e^{\texttt{sim}_f(t_i,t_j)}}{\sum_{n=1}^{N} e^{\texttt{sim}_f(t_n,t_j)}}\right) \end{aligned} \quad (7)$$

$$L = \lambda \cdot L_{cross} + L_{mono} \quad (8)$$

4 Experimental Settings

This section describes the details of our experimental setting, for both training and evaluation.

4.1 Training setup

The translation pairs used for training are downloaded from OPUS[1] (Tiedemann, 2012), where the volume of each language corpus is given in Appendix B. We focus on five languages: English (`en`), French (`fr`), Japanese (`ja`), Korean (`ko`), and Russian (`ru`). We train two types of models, cross-lingual and multi-lingual. For each cross-lingual model, we use `en-ko`, `en-ja`, `en-ru`, and `en-fr` pairs, respectively. For the multi-lingual model, we train with all translation pairs for five languages.

As discussed in Sec. 3, we consider two types of soft label $w(i,j)$: the Priority label in Eq. 4 defines the soft label by using the similarity measured at

[1] https://opus.nlpl.eu

a mono-lingual embedding (for a pre-defined anchor language), while the Average label in Eq. 5 uses the similarity averaged out over mono-lingual embeddings of both source and target language. Note that we need to choose the anchor language (among the source and the target language), for the former one. By default, we set the priority of the languages based on the volume of each language corpus used in training, thus having the following order: en, ru, ja, fr, and ko. The anchor language is defined as the one with higher priority between language pair. We also apply TMS, which is shown in Eq. 8, using shared $w(i, j)$ for the monolingual alignment and cross-lingual alignment.

Each model is trained for 30 epochs[2] on 2 RTX-3090 GPUs with global batch size 32. We use the cross-accelerator to expand negative samples, as described in Appendix C.2. The initial learning rate is set to $\gamma = 5 \cdot 10^{-3}$, and we linearly decay the learning rate. We use the AdamW optimizer. We tune the temperature parameter on en-ko bilingual dataset, and set it to $\tau = 0.1$. Also, we set the portion of cross-lingual loss in TMS as $\lambda = 0.1$. We apply the mixed precision training, to improve the training efficiency.

4.2 Evaluation tasks

Bitext Mining We evaluate our model on three bitext mining datasets, Tatoeba (Artetxe and Schwenk, 2019), BUCC (Zweigenbaum et al., 2017) and FLORES-200 (Costa-jussà et al., 2022). Tatoeba and BUCC are English-centric translation pair benchmark datasets that are included in MTEB (Muennighoff et al., 2022), and FLORES-200 is a N-way parallel benchmark dataset. Throughout the paper, we use the average accuracy measured from both directions (e.g., en→ko and ko→en) for BUCC and Tatoeba. We measure the average xSIM error rate from (Heffernan et al., 2022) for each languages in FLORES-200.

Semantic Textual Similarity (STS) We evaluate our model on STS datasets to examine how well mono-lingual and cross-lingual spaces are formed. We test on STS12-STS22 and the STS benchmark in MTEB (Muennighoff et al., 2022), and measure the average spearman correlation for each of the en, ko, fr, ru and en-fr.

[2]We early stopped with Tatoeba validation set. Most of the trains were stopped at between 10 and 20 epoch.

Lang	Student Model	Teacher Model	Tatoeba (en-xx)	BUCC (en-xx)	STS (en)	STS (xx)
en-ko	$mE5_{base}$	$mE5_{base}$	**0.917**	-	**0.777**	**0.762**
		$E5_{base}$	0.907	-	0.759	0.740
		MPNet	0.869	-	0.692	0.685
	XLM-R	$mE5_{base}$	0.896	-	0.704	0.707
		$E5_{base}$	0.897	-	0.702	0.702
		MPNet	0.864	-	0.648	0.650
en-fr	$mE5_{base}$	$mE5_{base}$	**0.963**	**0.982**	**0.783**	0.775
		$E5_{base}$	0.956	0.973	0.764	0.782
		MPNet	0.944	0.963	0.706	**0.785**
	XLM-R	$mE5_{base}$	0.951	0.973	0.699	0.744
		$E5_{base}$	0.949	0.961	0.692	0.747
		MPNet	0.942	0.956	0.637	0.761

Table 1: Comparison of various combinations of student and teacher models, in terms of the bitext mining (accuracy) and STS (spearman correlation score) performances. The best performance is achieved when both teacher and student use $mE5_{base}$ model.

5 Results

We first test the model trained with a single language pair, and then show the result when the model is trained with multiple language pairs.

5.1 Effect of the Student/Teacher Model

Table 1 shows the effect of the (student, teacher) model pair on the performance of our soft contrastive loss, using loss in Eq. 6, without TMS. We test two student model architectures, $mE5_{base}$ (Wang et al., 2024) and XLM-R (Conneau et al., 2019), and three teacher models, $mE5_{base}$ (Wang et al., 2024), $E5_{base}$ (Wang et al., 2022), and MPNet[3] (Song et al., 2020). The details of the student model selection are described in Appendix C.1. One can confirm that using $mE5_{base}$ for both teacher and student performs the best in both STS and bitext mining tasks. Thus, for the following experiments, we use $mE5_{base}$ for both teacher and student as a baseline.

5.2 Effectiveness of Our Loss

Recall that we propose training a student model using the contrastive loss with soft labels obtained from the teacher model. We denote our method as *soft contrastive loss*, and observe the effect of different loss functions in Table 2 and Table 3. Given a pre-trained student model, we fine-tune it with different losses. We compare two types of contrastive loss in Eq. 2, where one uses *soft* label $w(i, j)$ (Priority label in Eq. 4) with TMS (Eq. 8) and the other uses *hard* label $w(i, j)$ in Eq. 3. Note that mUSE (Yang et al., 2019) and LaBSE (Feng et al., 2020) use hard labels in contrastive loss where the translation pair is the only available positive pair,

[3]https://huggingface.co/sentence-transformers/all-mpnet-base-v2

Lang	Loss	Tatoeba (en-xx)	BUCC	STS (en)	STS (xx)
en-ko	Soft Contrastive (Ours)	**0.916**	-	0.788	0.778
	Hard Contrastive	0.863	-	0.674	0.675
	MSE (Reimers and Gurevych, 2020)	0.911	-	**0.803**	**0.793**
	mUSE (Yang et al., 2019)	0.853	-	0.715	0.698
	Pretrained Model	0.873	-	0.802	0.777
en-fr	Soft Contrastive (Ours)	**0.960**	**0.987**	0.796	**0.791**
	Hard Contrastive	0.937	0.933	0.675	0.748
	MSE (Reimers and Gurevych, 2020)	0.959	0.980	**0.803**	0.704
	mUSE (Yang et al., 2019)	0.950	0.984	0.713	0.771
	Pretrained Model	0.951	0.984	0.802	0.781
en-ja	Soft Contrastive (Ours)	**0.956**	-	0.798	-
	Hard Contrastive	0.933	-	0.730	-
	MSE (Reimers and Gurevych, 2020)	0.949	-	**0.808**	-
	mUSE (Yang et al., 2019)	0.925	-	0.742	-
	Pretrained Model	0.931	-	0.802	-
en-ru	Soft Contrastive (Ours)	**0.951**	**0.979**	0.787	**0.616**
	Hard Contrastive	0.949	0.955	0.666	0.545
	MSE (Reimers and Gurevych, 2020)	0.945	0.978	**0.804**	0.601
	mUSE (Yang et al., 2019)	0.944	0.978	0.720	0.548
	Pretrained Model	0.936	0.978	0.802	0.615

Table 2: Comparison of different loss functions used for fine-tuning pre-trained student model, tested on bitext mining tasks and STS tasks. The gray shaded method is the baseline which uses the pre-trained student model as it is. The best performance is indicated in bold, second most performance is indicated with an underline, throughout this paper. For Tatoeba dataset, our *soft* contrastive loss outperforms all compared losses.

Pretrained model	Fine-tune loss	Tatoeba	BUCC	FLORES-200
$mE5_{base}$	Soft Contrastive (Ours)	**0.949**	**0.983**	**0.02**
$mE5_{base}$	MSE	0.942	0.975	0.05
$mE5_{base}$	-	0.923	0.981	0.16
MPNet-multilingual	-	0.945	0.970	0.28
LASER2	-	0.939	0.981	0.20
LaBSE	-	**0.948**	**0.985**	**0.01**

Table 3: Comparison between existing models and fine-tuning loss on multi-lingual data, tested on bitext mining. Measured accuracy for Tatoeba and BUCC, while measuring xSIM error rate for FLORES-200. Note that fine-tuned with MSE is the same approach as Reimers and Gurevych (2020). Ours show similar performance to current SoTA, LaBSE.

corresponds to Eq. 3. We also test using MSE loss for distilling the embeddings of the teacher model to the embeddings of the student model, as in (Reimers and Gurevych, 2020).

Table 2 provides the performances tested on bitext mining tasks and STS tasks trained with a single language pair, i.e. cross-lingual version of ours. We test on four different language pairs {en-xx} where xx is either ko, fr, ja, or ru.

We have three major observations. First, our *soft* contrastive loss outperforms conventional *hard* contrastive loss in all performance metrics in all language pairs. For example, in Tatoeba dataset, our method has up to 5.3% accuracy gain (*e.g.,* from 86.3% to 91.6% for en-ko pair) compared with hard contrastive loss. Note that compared with the pre-trained model (shown in the gray shaded region in Table 2), additional training with hard contrastive loss sometimes harms the performance, *e.g.,* the accuracy degrades from 87.3% to 86.3% in Tatoeba dataset for the model trained with en-ko pair, and the STS performance degrades from 0.802 to 0.666 for the model trained with en-ru pair, which is critical.

Second, our *soft* contrastive loss provides the best performance in the bitext mining task, Tatoeba, and BUCC, for all language pairs. Compared with the pre-trained student model, additional training with soft contrastive loss improves the accuracy up to 4.3%.

Third, the STS performance for non-English languages is improved, after training with our soft contrastive loss. For example, after training with en-fr translation pair, the STS performance elevates by 0.01 when using soft contrastive loss, while 0.077 degradations (from 0.781 to 0.704) shown in MSE loss (Reimers and Gurevych, 2020).

Furthermore, we demonstrate the effectiveness of our loss through training with multiple language pairs, i.e. multi-lingual version of ours. Table 3 shows the bitext mining performances of multi-lingual models, tested on five languages, en, ko, ja, fr and ru. We also test the performance of pre-trained multi-lingual model checkpoints, namely, $mE5_{base}$ (Wang et al., 2024), LASER2 (Artetxe and Schwenk, 2019), LaBSE (Feng et al., 2020), and MPNet-multilingual[4] (Reimers and Gurevych, 2020). We trained with en-xx pairs (en-ko, en-ja, en-ru, and en-fr) for the fine-tuning. As a result, ours outperforms Reimers and Gurevych (2020) in every bitext mining task. Moreover, compared to other pretrained models, ours shows close results to current bitext mining State-of-the-Art, LaBSE.

5.3 Factor Analysis on Our Method

Priority vs Average We compare the two soft label methods we proposed in Eq. 4, 5 by varying the label functions, shown in Table 4, 5. *Priority* and *Average* stand for the loss described in Eq. 6 with $w(i, j)$ from Eq. 4, 5, respectively. TMS stands for the loss with monolingual alignment shown in Eq. 8.

Table 4, 5 shows the performance of model trained with single language pair data and multiple language pairs data respectively (cross-lingual and multi-lingual). Both *Priority* and *Average* significantly improved the performance of most bitext mining compared to the pretrained model. While

[4] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

Lang	Loss	Tatoeba (en-xx)	BUCC	STS (en)	STS (xx)
en-ko	*Average*	0.912	-	0.771	0.757
	Priority	**0.917**	-	0.777	0.762
	Priority + TMS	0.916	-	0.788	**0.778**
	Pretrained Model	0.873	-	**0.802**	0.777
en-fr	*Average*	0.959	0.981	0.767	**0.794**
	Priority	**0.963**	0.982	0.783	0.775
	Priority + TMS	0.960	**0.987**	0.796	0.791
	Pretrained Model	0.951	0.984	**0.802**	0.781
en-ja	*Average*	0.957	-	0.787	-
	Priority	**0.960**	-	0.787	-
	Priority + TMS	0.956	-	0.798	-
	Pretrained Model	0.931	-	**0.802**	-
en-ru	*Average*	**0.955**	**0.980**	0.775	0.612
	Priority	0.953	0.979	0.781	0.607
	Priority + TMS	0.951	0.979	0.787	**0.616**
	Pretrained Model	0.936	0.978	**0.802**	0.615

Table 4: Comparison of each variation trained with cross-lingual data, in terms of the bitext mining and STS performances. *Priority* shows slightly better performance than *Average* in bitext mining tasks, except for en-ru. Applying TMS enhances the STS performance, better than a pre-trained model for the non-English language.

Loss	Parallel Corpus	Tatoeba	BUCC	FLORES-200
Average	All pairs	0.950	0.979	0.04
Priority	All pairs	**0.952**	0.978	0.04
Priority + TMS	All pairs	0.948	0.983	0.04
Average	en-xx	0.948	0.979	0.04
Priority	en-xx	0.942	0.979	0.05
Priority + TMS	en-xx	0.949	**0.983**	**0.02**
Pretrained Model	-	0.923	0.981	0.16

Table 5: Comparison of different loss on multi-lingual data, in terms of the bitext mining task performances.

the performance gap between *Priority* and *Average* is trivial in Table 5, *Priority* shows slightly better performance than *Average* for bitext mining in Table 4. Yet, *Average* performs better on STS (xx) performances. For example, for models trained with en-fr, *Average* achieves 0.794, which 0.019 higher than the *Priority* (0.775)

Effect of TMS We validate the effectiveness of TMS by comparing *Priority* and *Priority* + TMS in Table 4, 5. Table 4 shows that using TMS improves the performance significantly on STS in all language pairs. For example, TMS increases STS performances with en-ko pairs from 0.011 higher on STS (en) and 0.016 higher on STS (xx). Though there was no performance gain in bitext mining after applying TMS, still *Priority* + TMS shows much better performance than the pre-trained model.

The impact of TMS is more dramatic in a multi-lingual experiment setting, shown in Table 5. Applying TMS shows the best STS performance shown in Appendix A, and even the best performance in bitext mining tasks shown in Table 5.

Effect of Language Pair Selection We expect there was an interference that arose from using multiple languages as a teacher, as there was less performance gain for BUCC and FLORES-200 when expanding a corpus size (from using only en-xx to all pairs). Thus, we made an additional experiment to analyze the effects of language selection on the performance.

Data	STS en	STS ko	STS fr	STS ru	STS en-fr	Tatoeba	BUCC	FLORES-200
en-xx, fr-xx	**0.757**	0.694	**0.742**	**0.589**	**0.765**	**0.948**	**0.983**	**0.04**
en-xx, ru-xx	**0.757**	**0.696**	0.718	0.586	0.758	0.946	0.979	0.07

Table 6: Comparison of varying language pairs for train corpus, tested on bitext mining tasks) and STS tasks. The model trained on en-xx, fr-xx performs better than the model trained on en-xx, ru-xx for STS in all languages.

Table 6 shows the results of training with our method on less language pairs. We test our method on language pairs containing en or fr (denoted by en-xx and fr-xx), and language pairs containing en or ru (denoted by en-xx and ru-xx). All tests in Table 6 are trained without TMS on the priority labels using our loss.

Using the pair en-xx, fr-xx performs better than using en-xx, ru-xx in most of the benchmarks. Not only bitext mining but also STS shows better performances for most languages. Even for ru STS, we can observe that en-xx, fr-xx performs better than en-xx, ru-xx. This can be seen as a synergy or interference between languages, which has a significant impact on performance. Thus, by selecting teacher languages that share similar monolingual embedding space, we believe we can achieve much better performance in multilingual tasks. We leave this as a future work.

6 Conclusion

In this paper, we proposed a method of improving multi-lingual embeddings, with the aid of the sentence similarity information measured at the mono-lingual teacher models. Our method can be considered as a variant of existing contrastive learning approach, where our method uses *soft* labels defined as the sentence similarity, while existing methods use *hard* labels. We tested our method on five different languages including en, ko, ja, fr, and ru. Our method shows the best performance in the Tatoeba dataset, and achieved high performance in other bitext mining tasks as well as STS tasks.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, pages 597–610.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning cross-lingual sentence representations via a multi-task dual-encoder model. *arXiv preprint arXiv:1810.12836*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv e-prints*, pages arXiv–2308.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Jiyeon Ham and Eun-Sol Kim. 2021. Semantic alignment with calibrated similarity for multilingual sentence embedding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1781–1791.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. *arXiv preprint arXiv:2205.12654*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2020. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J Passonneau. 2022. Contrastive data and learning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 39–47.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the second bucc shared task: Spotting parallel sentences in comparable corpora. In

Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 60–67.