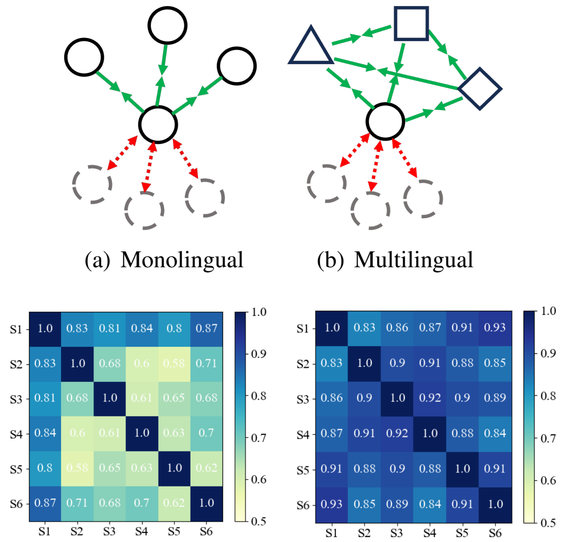
Leveraging Multilingual Positive Instances in Contrastive Learning to Improve Sentence Embedding

Kaiyan Zhao[*], **Qiyu Wu**[*], **Xin-Qiang Cai, Yoshimasa Tsuruoka**
The University of Tokyo, Tokyo, Japan
{zhaokaiyan1006, qiyuw, caixq, yoshimasa-tsuruoka}@g.ecc.u-tokyo.ac.jp

Abstract

Learning multilingual sentence embeddings is a fundamental task in natural language processing. Recent trends in learning both monolingual and multilingual sentence embeddings are mainly based on contrastive learning (CL) among an anchor, one positive, and multiple negative instances. In this work, we argue that leveraging multiple positives should be considered for multilingual sentence embeddings because (1) positives in a diverse set of languages can benefit cross-lingual learning, and (2) transitive similarity across multiple positives can provide reliable structural information for learning. In order to investigate the impact of multiple positives in CL, we propose a novel approach, named MPCL, to effectively utilize multiple positive instances to improve the learning of multilingual sentence embeddings. Experimental results on various backbone models and downstream tasks demonstrate that MPCL leads to better retrieval, semantic similarity, and classification performance compared to conventional CL. We also observe that in unseen languages, sentence embedding models trained on multiple positives show better cross-lingual transfer performance than models trained on a single positive instance.



(a) Monolingual (b) Multilingual

(c) Monolingual (d) Multilingual

Figure 1: Different shapes denote examples in different languages. Solid and dotted arrows denote positive and negative pairs, respectively. (a) vs. (b): Multilingual positives by translation exhibit transitive similarity, while monolingual positives do not. (c) vs. (d): Pairwise semantic similarity scores of sampled sentences for mono and multilingual, highlighting the similarity transitive similarity. Example sentences are sourced from the XNLI dataset, refer to A.1 for details.

1 Introduction

Multilingual sentence embedding transforms sentences in different languages into a shared embedding space (Feng et al., 2020; Wang et al., 2022b), where sentences with similar meanings are positioned close to each other. This is a fundamental and important task in Natural Language Processing (NLP), with various applications including multilingual retrieval (Yang et al., 2020), cross-lingual classifications (Hirota et al., 2020), and multilingual inference (Conneau et al., 2018).

As contrastive learning (CL) exhibits great strength on learning sentence representation, CL-based methods have become the common practice for learning monolingual (Gao et al., 2021; Su et al., 2022; Ni et al., 2022) as well as multilingual sentence embeddings (Feng et al., 2020; Wang et al., 2022b). Typically, conventional CL is performed with an anchor, a positive and multiple negative examples. The learning objective of CL is to pull the anchor and the positive closer and push the anchor and the negatives apart (van den Oord et al., 2018).

Our work aims at improving CL for multilingual sentence embedding with *multiple positives*. While existing approaches in multilingual sentence embedding only takes the naive CL with a single positive example, we argue that leveraging multiple

[*]Equal contribution.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics Volume 1: Long Papers, pages 976–991

positives should be considered especially for multilingual sentence embeddings. In contrast to monolingual CL, richer and more complex relationships exist among multiple positives in multilingual CL: **(1) positives in a diverse set of languages**, which can benefit cross-lingual learning; **(2) transitive similarity across multiple positives by translation**, which provides reliable structural information for learning.

To show such properties, we calculate similarity scores among multiple positives to emphasize the unique effect of multiple positives especially in multilingual scenarios. In the monolingual setting shown in Figure 1(a) and 1(c), although multiple positive examples all share high similarity with the anchor, the transitive similarity does not always exist among positives, e.g., (S1, S2) and (S1, S3) have similar meanings but (S2, S3) do not. By contrast, in multilingual settings shown in Figure 1(b) and 1(d) where translations are used as positives, multiple positives can provide cross-lingual information from diverse languages. Moreover, multilingual translation guarantees transitivity of similarity across positive examples, leading to effective CL with multiple positives.

Motivated by the aforementioned discussion, in this paper, we investigate the impact of CL with multiple positives especially for multilingual sentence embeddings and propose MPCL (**M**ultilingual **P**ositives in **C**ontrastive **L**earning), a novel approach for sentence embedding to effectively leverage multiple positives to improve the quality of multilingual sentence embeddings. Specifically, we construct multiple positive instances by collecting multilingual translations for the anchor sentence. Besides, we propose to utilize a multi-positive loss function to effectively learn from the multiple positives, in which the conventional contrastive correlation and structural information among multilingual translations are learned simultaneously. To the best of our knowledge, we are the first to explore the impact of multiple positives on multilingual sentence embeddings.

Extensive experiments on various models and downstream tasks are conducted to evaluate the proposed approach. Experimental results confirm that leveraging multiple positives leads to better semantic similarity, retrieval, and classification performance on LaBSE (Feng et al., 2020), with an improvement of 4.1 on the BUCC task, 2.7 on the STS17 task, and 1.5 on the MTOP domain classification task, respectively. The improvement holds for a diverse set of backbone models, including the continual training on well-trained sentence embedding models such as mSimCSE (Wang et al., 2022b), as well as training from scratch on pre-trained language models such as mBERT (Devlin et al., 2019a) and XLM-RoBERTa (Conneau et al., 2020). A variant of MPCL outperforms the state-of-the-art model, mSimCSE, in various evaluation tasks. We also observe better cross-lingual transfer performance on unseen languages with our proposed MPCL compared to conventional CL with a single positive instance. Moreover, to investigate the effectiveness of MPCL, we evaluate variants of MPCL by incorporating different languages and adjusting the number of languages included in our training dataset.

2 Multiple Positives in Contrastive Learning for Multilingual Sentence Embeddings

This section begins with a formal definition of learning sentence embeddings with CL, followed by the construction of training data with multiple positives and the utilization of multiple positives for sentence embeddings.

Contrastive learning for sentence embeddings Given a sentence $x_i \sim \mathbb{X}$, sentence embedding learning aims to learn a parameterized network f_θ. The network can be applied to x_i to obtain a dense vector, i.e., $\mathbf{h}_i = f_\theta(x_i) \in \mathbb{R}^d$, which can represent the semantic meaning of sentence x_i. The idea of contrastive learning is to construct a positive example x_i^+ for x_i and pull them close, while keeping x_i far from other negative examples. A commonly used training objective (van den Oord et al., 2018; Gao et al., 2021; Wu et al., 2022) is to minimize the following contrastive loss:

$$l_i^{(s)} = -\log \frac{e^{sim(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} e^{sim(\mathbf{h}_i, \mathbf{h}_j)/\tau}}, \tag{1}$$

where $\mathbf{h}_i^+ = f_\theta(x_i^+)$, $sim(\cdot, \cdot)$ is a similarity metric, N is the size of a mini-batch, and τ is the temperature parameter. After training, these semantically meaningful embeddings can be used to represent sentences for various downstream tasks such as sentence retrieval, sentence-level classification, and semantic textual similarity.

Multiple Positives in Contrastive Learning The conventional approach cannot fully capture

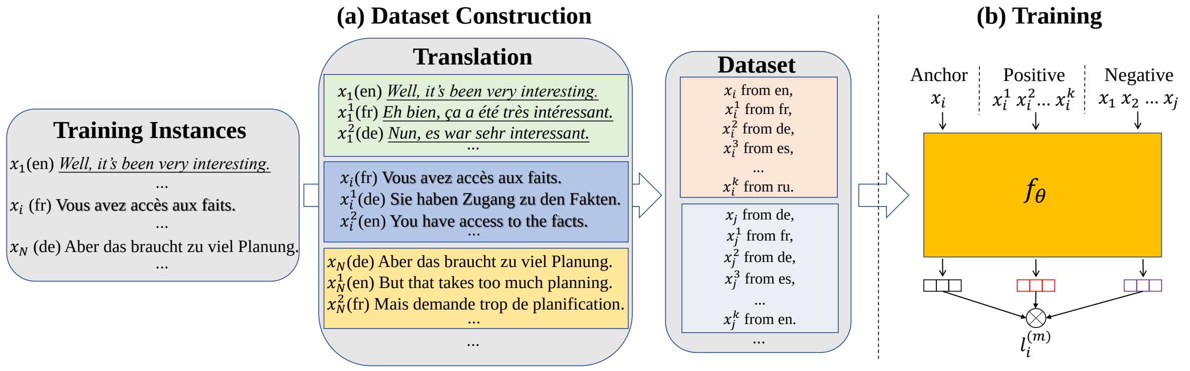


Figure 2: Illustration of MPCL. Left: we reorganize multilingual data with a translation dataset to construct a training dataset with multiple positives. Sentences in the same font are translations from different languages. Right: we perform contrastive loss with multiple positive instances to update the model.

the semantic richness and diverse expressions in different languages. To address this limitation, in this work, we propose to leverage multiple positives in contrastive learning to improve the learning of multilingual sentence embeddings. Unlike the conventional single-positive CL loss in Equation 1, for anchor sentence x_i, we construct a multiple positives set from multilingual translation $X_i^{mp} = \{x_i^1, ..., x_i^K\}$, where K is the number of positives from different languages. Inspired by previous methods that deal with multiple positives (Frosst et al., 2019; Khosla et al., 2020), the training objective of CL with multiple positives in multilingual sentence embedding is shown as:

$$l_i^{(m)} = -\log \frac{\sum_{k=1}^{K} e^{sim(\mathbf{h}_i, \mathbf{h}_i^k)/\tau}}{\sum_{j=1 \wedge j \neq i}^{N} e^{sim(\mathbf{h}_i, \mathbf{h}_j)/\tau}}, \quad (2)$$

where $\mathbf{h}_i^k$ stands for the representation of positive sentence x_i^k from positive set X_i^{mp}, and N is the size of a mini-batch. Equation 2 allows us to capture the linguistic diversity and complex relationships among sentences across different languages.

Dataset Construction in MPCL Figure 2 illustrates the dataset construction in MPCL. In order to utilize the transitivity among positives, we collect translations as multiple positives X_i^{mp} from a multilingual translation dataset for x_i. We reorganize the multilingual training instances by assembling translations into one group. Sentences within the same group share the same meaning and exhibit transitive similarity, allowing us to have one anchor sentence x_i and a positive set X_i^{mp} to perform the MPCL loss in Equation 2.

3 Experiments

3.1 Details of Training Dataset

The multilingual translation dataset used in our experiment is the XNLI (Conneau et al., 2018) dataset. Considering the intersection of different evaluation tasks, six languages, English (*en*), German (*de*), French (*fr*), Spanish (*es*), Russian (*ru*), and Chinese (*zh*), are selected in our dataset so that we can evaluate the effects on both seen and unseen languages simultaneously with minimal influence from other languages. Other combinations of languages will be discussed in Section 3.7. Sentences in languages other than English are derived from translations given in XNLI. This allows us to assemble multilingual translations into the same group. Specifically, when dealing with a given sentence, we exclusively choose the sentence itself, omitting its corresponding entailment, neutral, and contradictory counterparts provided in XNLI.

Finally, our dataset comprises 400k data groups. In this dataset, for each anchor sentence, we can access multiple positives at the same time. Note that each language has the same probability of serving as the anchor sentence. We perform a wide range of experiments with various models on this dataset to verify the effects of multiple positives.

3.2 Baselines

Several strong baselines are chosen for comparison. The first selections are two state-of-the-art multilingual sentence embedding models trained on one single positive instance, LaBSE (Feng et al., 2020) and mSimCSE (Wang et al., 2022b). We specifically choose $mSimCSE_{all}$, a variant of mSimCSE that includes 15 languages during training and utilizes hard-negative examples.

In addition, we select some general models such as Sentence-T5 (Ni et al., 2022), LASER, and LASER2 (Artetxe and Schwenk, 2019) to compare performance on different tasks. Besides, `bert-base-multilingual-uncased`[1] (Devlin et al., 2019b) and `xlm-roberta-large`[2] (Conneau et al., 2020) are included as alternative Pretrained Language Models (PLMs).

3.3 Training Details

We continuously train various backbone models including LM base models and sentence embedding base models. The training of all our models is conducted on one NVIDIA A100 80G. The batch size is set to 128, the maximum sequence length is set to 64, and the learning rate is 1e-5. Particularly for the LM base models mBERT and XLM-RoBERTa, the conventional contrastive loss in Equation 1 is initially used for warm-up, during which the learning rate is set to 2e-5 for 2000 steps.

The temperature parameter τ is set to 0.05. We use the `[cls]` token as sentence embedding. Cosine similarity is used as the similarity metric, which allows us to compute the similarity distribution by contrasting the anchor sentence with multiple positives and negatives. Accounting for the margin among multiple positives and negative examples, we specifically use min-max scaling to rescale the similarity scores within a range of $[-1/\tau, 1/\tau]$. Under our specified training details, the BERT-base size models require approximately 30G of memory, while the BERT-large size models require about 60G of memory.

We evaluate the models on development sets every 125 steps to find the best checkpoints. Specifically, STS22 and STS17 are used as the development set for each other. We also use Tatoeba and BUCC as each other's development sets for bi-text mining tasks. For classification tasks, we directly use the validation set provided by MTOP domain classification as the development set. All of our results are obtained from an average of five random seeds.

3.4 Evaluation Tasks

We evaluate models on three fundamental multi/cross-lingual tasks: bitext mining, semantic similarity and classification. We run semantic similarity and classification with MTEB[3] (Muennighoff et al., 2022), and bitext-mining with XTREME[4] (Hu et al., 2020) benchmark.

Bitext Mining is a retrieval task where a sentence and a paragraph (or longer sentence) will be given. The tested model is supposed to find the best match for the sentence in the paragraph by calculating cosine similarity for each pair of embedded sentences. We evaluate our trained models using 14 and 36 Tatoeba (Artetxe and Schwenk, 2019) and BUCC (Zweigenbaum et al., 2017) datasets through the XTREME benchmark. We report the F1 score for BUCC and the accuracy for Tatoeba.

Semantic Similarity requires the models to calculate the similarity (scores) of two given sentences. Higher scores generally mean higher similarity. We choose the cross-lingual STS17 (Cer et al., 2017) and STS22 (Chen et al., 2022) and report Spearman correlation scores based on cosine similarity metrics. Note that in STS22, all of our averaged results do not contain French-Polish (*fr-pl*) because we find this pair in the MTEB benchmark appears to be unstable, and even totally different models can have exactly the same correlation score on the MTEB leaderboard.

Classification tasks require the model to determine the label of given sentences based on their sentence embeddings. An additional classifier layer will be trained on the given training set, and the performance of the model will be tested on the test set. We choose the MTOP Domain Classification task (Li et al., 2021a) through the MTEB benchmark and report the accuracy metric.

3.5 Main Experimental Results

In this section, we present the main experimental results. In particular, + Multiple refers to models that are trained through our proposed framework with five positive instances. To facilitate a fair comparison with conventional CL with one single positive, we modify our dataset to follow a parallel structure, where only source-target pairs from different languages are included. For example, in our main experiments, we have six multilingual sentences in one group so this group will be converted into three random language pairs in the parallel dataset. + Single refers to the models that are trained on

[1] https://huggingface.co/bert-base-multilingual-uncased

[2] https://huggingface.co/xlm-roberta-large

[3] https://huggingface.co/spaces/mteb/leaderboard

[4] https://github.com/google-research/xtreme

Model	BUCC	Tatoeba(14avg.)	Tatoeba(36avg.)	STS17	STS22	MTOP	Avg.[†*]
LM Base Models							
mBERT (Devlin et al., 2019b)	56.7	-	-	-	-	-	-
+ Single	$84.1_{\pm 0.09}$	$70.5_{\pm 0.34}$	$64.4_{\pm 0.45}$	$57.0_{\pm 0.26}$	$53.4_{\pm 0.63}$	$62.3_{\pm 0.21}$	65.3
+ Multiple (Ours)	$\mathbf{85.3}_{\pm 0.36}$ (↑**1.2**)	$\mathbf{71.6}_{\pm 0.20}$ (↑**1.1**)	$\mathbf{65.1}_{\pm 0.34}$ (↑**0.7**)	$\mathbf{57.8}_{\pm 0.31}$ (↑**0.8**)	$\mathbf{55.8}_{\pm 0.56}$ (↑**2.4**)	$\mathbf{62.7}_{\pm 0.20}$ (↑**0.4**)	66.4
XLM-R (Conneau et al., 2020)	66.0	57.6	53.4	-	-	-	-
+ Single	$94.5_{\pm 0.17}$	$91.3_{\pm 0.08}$	$89.6_{\pm 0.19}$	$71.1_{\pm 0.63}$	$59.8_{\pm 0.44}$	$83.0_{\pm 0.19}$	81.6
+ Multiple (Ours)	$\mathbf{95.7}_{\pm 0.36}$ (↑**1.2**)	$\mathbf{92.0}_{\pm 0.11}$ (↑**0.8**)	$\mathbf{90.4}_{\pm 0.13}$ (↑**0.8**)	$\mathbf{73.2}_{\pm 0.20}$ (↑**2.1**)	$\mathbf{61.4}_{\pm 0.41}$ (↑**1.6**)	$\mathbf{84.5}_{\pm 0.39}$ (↑**1.5**)	82.9
XLM-R w/ hard negative	-	-	-	-	-	-	-
+ Single (mSimCSE$_{all}$) (Wang et al., 2022b)	95.2	93.2	91.4	76.7	63.2	84.1	84.0
+ Multiple (Ours)[†]	$\mathbf{95.4}_{\pm 0.27}$ (↑**0.2**)	$\mathbf{93.5}_{\pm 0.13}$ (↑**0.3**)	$\mathbf{91.8}_{\pm 0.07}$ (↑**0.4**)	$\mathbf{78.9}_{\pm 0.47}$ (↑**2.2**)	$\mathbf{64.0}_{\pm 0.16}$ (↑**0.8**)	$\mathbf{86.8}_{\pm 0.30}$ (↑**2.7**)	85.1
Sentence Embedding Base Models							
INFOXLM (Chi et al., 2021)	-	77.8	67.3	-	-	-	-
LASER (Artetxe and Schwenk, 2019)	92.9	95.3	84.4	-	-	-	-
LASER2 (Artetxe and Schwenk, 2019)	-	-	-	69.2	41.6	73.5	-
Sentence-T5-large (Ni et al., 2022)	-	-	-	44.4	47.0	61.5	-
mSimCSE$_{all}$ (Wang et al., 2022b)	95.2	93.2	91.4	76.7	63.2	84.1	84.0
+ Multiple (Ours)	$96.0_{\pm 0.23}$ (↑0.8)	$93.5_{\pm 0.10}$ (↑0.3)	$91.4_{\pm 0.14}$	$\mathbf{78.5}_{\pm 0.23}$ (↑**1.8**)	$\mathbf{64.3}_{\pm 0.28}$ (↑**1.1**)	$85.9_{\pm 0.27}$ (↑1.8)	84.9
LaBSE (Feng et al., 2020)	93.5	95.3	95.0	74.2	60.9	84.6	83.9
+ Multiple (Ours)	$\mathbf{97.6}_{\pm 0.19}$ (↑**4.1**)	$\mathbf{96.0}_{\pm 0.08}$ (↑**0.7**)	$\mathbf{95.4}_{\pm 0.07}$ (↑**0.4**)	$76.9_{\pm 0.22}$ (↑2.7)	$61.5_{\pm 0.29}$ (↑0.6)	$\mathbf{86.1}_{\pm 0.26}$ (↑**1.5**)	85.6

Table 1: Overall results of different models on various downstream tasks. We report the average scores and their corresponding standard deviation from five random seeds for each task. We adopt the baseline's results from Hu et al. (2020) and Muennighoff et al. (2022). +Single stands for models we continually train on parallelized dataset with one single positive while +Multiple stands for models we train on our proposed method with multiple positives. [†]: This variant is trained with all 15 languages in XNLI and combined with hard negatives. [†*]: This column refers to the average score of all six tasks, showing statistically significant results with p-value < 0.005 when comparing each +Multiple to its corresponding base model.

this modified dataset. Note that the data in this modified dataset are the same as those in the original dataset, with the only difference being their structure, and the model has an equal chance of seeing each sentence once in both datasets. Unless otherwise specified, both + Multiple and + Single refer to the models trained without hard negatives.

More specifically, we aim to address two questions based on the overall results of various downstream tasks shown in Table 1. **Q1**: Does the utilization of multiple positives yield more substantial benefits to the model compared to conventional CL with a single positive? **Q2**: Does the effectiveness of leveraging multiple positives still hold for stronger sentence embedding models? Refer to Appendix A.2 for detailed results of different models.

3.5.1 Multiple Positives Yield Better Performance than Single Positive

We first train general LMs with a warm-up under our proposed methods and compare their performance with LMs trained on conventional CL.

From the upper part of Table 1, it is evident that continuous training with both single and multiple positives on pretrained language models enhances the performance of the models. However, from the average score of the downstream tasks, we can observe that models trained using our proposed framework with multiple positives demonstrate stronger performance than conventional CL-based methods with single positives. Specifically, we discover an average improvement of 1.1 for mBERT and 1.3 for XLM-R across different downstream tasks compared to models trained on single positive instance. The most significant improvement is seen in STS22 for mBERT, and in STS17 for XLM-R.

In order to fairly compare MPCL with state-of-the-art models, we train a variant with the same language coverage and apply hard negatives (Kalantidis et al., 2020). We include all 15 languages in XNLI and utilize sentences with contradictory labels as hard negatives. This comparison is referred to as **XLM-R w/ hard negative** shown in Table 1. Our XLM-R trained with multiple positives surpasses mSimCSE$_all$, which is trained with single positives in all tasks we report. Note that during the training of this variant, an anchor sentence can access five positives from different languages and the hard negative instance is also randomly selected from all languages. This variant indicates that our proposed MPCL can be incorporated with other existing orthogonal methods, such as hard negatives, and contributes to better performance.

These observations suggest that for general multilingual language models, leveraging multiple positives can offer a richer and more useful source of information for training, thus yielding more substantial benefits to the model. Note that with the exception of the BUCC task, all the other evaluation tasks include languages that are excluded from our dataset.

3.5.2 Multiple Positives Can Further Improve Sentence Embedding Models

Next, we continually train sentence embedding models, including LaBSE and mSimCSE to validate the effectiveness of multiple positives for pretrained sentence embedding models. For mSimCSE$_{all}$, their results of BUCC and Tatoeba are adopted from Wang et al. (2022b) while we use its released checkpoint [5] to evaluate its performance on STS17, STS22, and MTOP Domain Classification. LaBSE's results are adopted from the MTEB leaderboard.

Upon examining the lower part of Table 1, we can observe that the highest scores for each task are obtained by models continually trained on our framework. At the same time, it is evident that the improvements in the BUCC, STS, and MTOP Domain Classification tasks are more prominent compared to the Tatoeba dataset. A possible reason for this observation is that we only include five languages, excluding English in our dataset, in contrast to the comprehensive evaluation of 36 languages in the Tatoeba dataset. These findings imply that the utilization of multiple positives can even significantly enhance the performance of well-trained sentence embedding models, indicating the robustness of our proposed method.

3.6 Transferring to Unseen Languages

Despite the presence of unseen languages in all of our downstream evaluation tasks, models trained using multiple positives consistently demonstrate improvements across almost all overall results. This motivates us to delve deeper into the transferability of multiple positives. We show the results of all language pairs from STS17, STS22 and MTOP Domain Classification in Figure 3.

The value of 0 on horizontal axis stands for the scores of the origianl LaBSE and mSimCSE$_{all}$. For example, in Figure 3, the first pair English-German (*en-de*) indicates that parallel training with a single positive leads to a slight decrease in performance for both LaBSE and mSimCSE, while training with multiple positives leads to an improvement. Note that our training dataset only contains *en*, *de*, *fr*, *es*, *ru* and *zh* so all the other language pairs are considered unseen. Language pairs such as English-Turkish (*en-tr*), where only one of the two languages is observed in the training set are also considered as exclusive.

[5] https://github.com/yaushian/mSimCSE

In Figure 3, we can observe that models trained with multiple positives generally exhibit better transfer ability to unseen languages. For example, when looking at the unseen language pair *pl-en* in STS22, we observe a drop in accuracy with single positive training, while multiple positive training still shows a significant improvement in LaBSE. This trend can also be observed in mSimCSE's variance, indicating that the transfer ability of multiple positives surpasses that of single positive training. Average results of unseen languages in STS tasks can be found in Appendix A.2, Table 7.

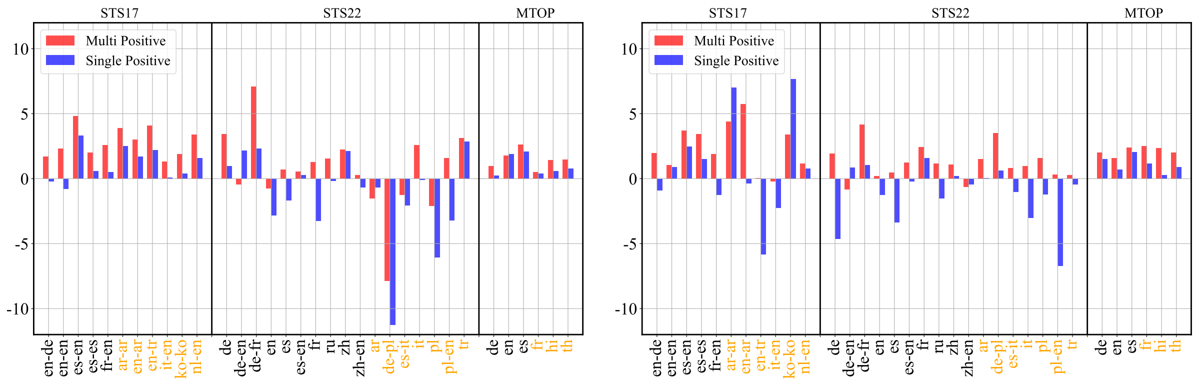
3.7 The Choice of Languages in Training Set

In our main experiments, we choose six different languages to satisfy various downstream tasks. However, the composition of the dataset, including the number and selection of positives, still remains unexplored. In this section, we alter the composition of training datasets and specifically choose XLM-R to assess different downstream tasks.

3.7.1 Training with Only Unseen Language

We first construct another training dataset using multiple positives extracted from XNLI which comprises only four languages: Bulgarian (*bg*), Greek (*el*), Vietnamese (*vi*) and Swahili (*sw*) that do not overlap with any language that will be tested in STS17 and STS22 so that all results exhibit models' fully transfer abilities. Besides, we add two more languages: Hindi (*hi*) and Thai (*th*) to see whether cross-lingual signals from more non-overlap languages can help improve the transfer ability. Figure 4 shows the results.

As reported in Wang et al. (2022b), we also observe that CL-based methods exhibit remarkable transfer abilities on totally unseen languages. From Figure 4, we can observe a slight drop when incorporating two additional languages into the dataset. This finding aligns with the observation presented by Conneau et al. (2020), where they highlight a trade-off between the number of languages and transfer performance. However, the strong transfer ability of multiple positives remains evident as training on completely non-overlapping languages can still yield competitive results compared to our original training dataset. Besides, we also observe an obvious trend in Figure 4 where all average results from multiple positives consistently surpass those from single positives. We believe that by bringing anchor sentences and all the remaining positives closer together, models can effectively



(a) Performance variation on LaBSE

(b) Performance variation on mSimCSE$_{all}$

Figure 3: Detailed performance changes of LaBSE and mSimCSE$_{all}$ on different languages after training on multiple positive or single positive. On the horizontal axis, 0 represents the original scores without further training. The vertical axis shows changes in performance after further training. Bars above the horizontal axis indicate improvements, while those below indicate decreases. The bold lines split the results into three different parts: STS17, STS22 and MTOP Domain Classification from left to right. Orange color highlights unseen languages.

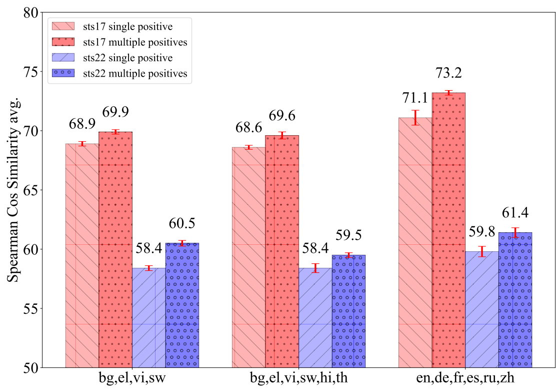


Figure 4: Results of the STS17 and STS22 tasks trained on fully none-overlapping languages. We report the average scores of all language pairs included in STS17 and STS22. Error bars refer to standard deviation.

capture the nuances of different languages and achieve better language representation.

3.7.2 Training with Only Seen Language

We also explore the effect of training with only languages that will be tested in the STS17 and STS22 tasks. We remove the *ru* and *zh* from our dataset, which will not be evaluated in STS17 to see whether *ru* and *zh* provide more useful cross-lingual signals. Besides, we add two more additional languages, Arabic (*ar*), and Turkish (*tr*) so that all the languages in this dataset (*en*, *de*, *fr*, *es*, *ar*, *tr*) will be tested in the STS17 and STS22 tasks. The results are shown in Table 2.

We focus specifically on the language pairs that we removed or added such as English-Arabic (*en-ar*) and English-Turkish (*en-tr*) to examine the impact of different dataset compositions. Upon removing *ru* and *zh* from the original dataset, we observe a slight decrease in the overall accuracy of STS17. This suggests that *ru* and *zh* in fact contribute valuable cross-lingual signals to the model in STS17. However, in STS22, although the performance of *ru* drops due to its removal, the accuracy of *zh* acquires an interesting improvement. One possible reason for this may be that although we include *zh* in our dataset, we do not have the exact same *zh-zh* pairs that are tested in STS22. Thus, the model is able to learn information from other cross-lingual signals.

In an attempt to replace *ru* and *zh* with *ar* and *tr*, we observe significant improvements in the *ar* and *tr* related pairs in both STS17 and STS22, but this change leads to a decline in the overall accuracy compared to our original results. Dhamecha et al. (2021) have noted that languages with high relatedness can mutually benefit each other. Therefore, adding languages like *ar* which has low linguistic relatedness with other languages may have an impact on the performance of other languages.

3.8 Case study

In this section, we randomly choose two examples from XNLI test set to demonstrate the effect of multiple positives on cross-lingual similarity. With the six languages included in our experiment, we can obtain fifteen cross-lingual pairs. The similarities are calculated for all fifteen language pairs. The re-

Model	en-ar	en-tr	STS17 avg.	ar	ru	tr	zh	zh-en	STS22 avg.
XLM-R									
+Single	$67.9_{\pm0.49}$	$68.4_{\pm0.71}$	$71.1_{\pm0.63}$	$57.2_{\pm0.64}$	$59.1_{\pm0.39}$	$64.4_{\pm0.43}$	$65.6_{\pm0.25}$	$67.4_{\pm0.65}$	$59.8_{\pm0.44}$
+Multiple	$71.0_{\pm0.24}$	$72.5_{\pm0.40}$	$\mathbf{73.2}_{\pm0.20}$	$57.1_{\pm0.48}$	$\mathbf{60.3}_{\pm0.20}$	$64.3_{\pm0.23}$	$64.8_{\pm0.27}$	$\mathbf{68.6}_{\pm0.22}$	$\mathbf{61.4}_{\pm0.41}$
+Single, *-ru, zh*	$67.6_{\pm0.18}$	$70.3_{\pm0.88}$	$70.6_{\pm0.64}$	$55.6_{\pm0.46}$	$57.5_{\pm0.28}$	$63.1_{\pm0.60}$	$65.3_{\pm0.17}$	$67.3_{\pm0.42}$	$59.6_{\pm0.46}$
+Multiple, *-ru, zh*	$70.3_{\pm0.49}$	$70.8_{\pm0.23}$	$71.8_{\pm0.48}$	$56.2_{\pm0.07}$	$58.4_{\pm0.67}$	$62.7_{\pm0.60}$	$65.7_{\pm0.28}$	$68.1_{\pm0.32}$	$60.2_{\pm0.31}$
+Single, *+ar, tr*	$71.2_{\pm0.61}$	$72.1_{\pm0.58}$	$71.3_{\pm0.44}$	$58.7_{\pm0.33}$	$58.3_{\pm0.32}$	$\mathbf{65.7}_{\pm0.26}$	$64.8_{\pm0.51}$	$67.0_{\pm0.37}$	$60.3_{\pm0.36}$
+Multiple, *+ar, tr*	$\mathbf{72.5}_{\pm0.18}$	$\mathbf{73.6}_{\pm0.49}$	$72.1_{\pm0.27}$	$\mathbf{59.0}_{\pm0.27}$	$59.3_{\pm0.61}$	$65.3_{\pm0.41}$	$\mathbf{66.3}_{\pm0.34}$	$67.3_{\pm0.35}$	$60.9_{\pm0.34}$

Table 2: Results of different composition of dataset on STS17 and STS22 task. *-ru, zh* means we only have four languages in the dataset while *+ar, tr* means that we replace *ru* and *zh* with *ar* and *tr*.

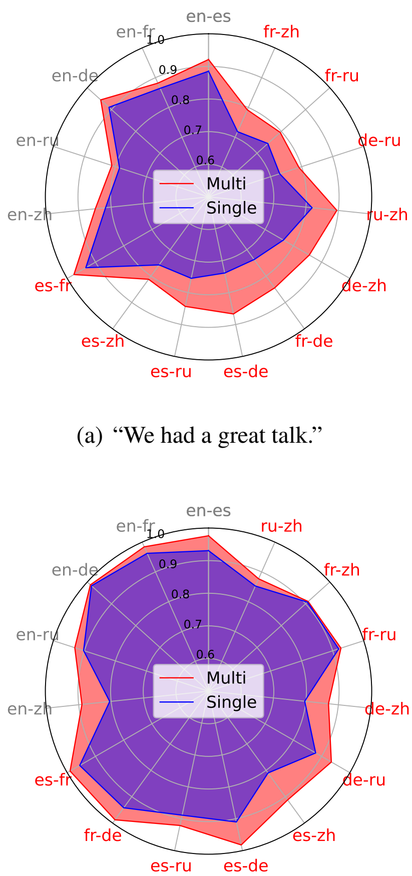
sults are shown in Figure 5. The sub-figure caption is the example sentence in English. Non-English-centric pairs are highlighted with red color. As we calculate similarity scores based on translations, the gold label similarity score for all cross-lingual pairs should be 1.0. As shown in the figure, it can be observed that for most English-centric pairs, single positive and multiple positive models achieve a comparable level of similarity. However, for non-English-centric pairs, multiple positive model exhibits an obvious higher similarity. This indicates that our approach of utilizing translations as multiple positives improves the cross-lingual representation learning, especially for non-English-centric language pairs.

4 Related Work

4.1 Contrastive Learning

Conventional contrastive learning performs with an anchor, one positive instance and multiple negative instances by pulling the distance between positive instances closer and between negative instances farther. The idea of contrastive loss can be traced back to Chopra et al. (2005). Later, the wide usage of contrastive loss in the field of computer vision takes it to a higher level and has been proven to be an effective way of learning representations (Wu et al., 2018; He et al., 2019; Chen et al., 2020). In the field of NLP, contrastive learning has been applied into a variety of tasks such as machine translation (Pan et al., 2021), text classification (Du et al., 2021), summarization (Duan et al., 2019). Recently, it has also been shown that CL plays a significant role in cross-modal representation learning (Li et al., 2021b; Radford et al., 2021) which indicates that even pulling positive instances from different modalities can be beneficial.

Contrastive learning with multiple positives has been studied in previous researches in computer vision (Khosla et al., 2020), with fine-grained strategies such as soft-nearest neighbor (Frosst et al., 2019) and ranking (Dwibedi et al., 2021; Hoffmann et al., 2022). In this paper, as the focus is verifying the impact of the usage of multiple translated positives for sentence embedding, we simply assign equal importance to all positives.



(a) "We had a great talk."

(b) "I don't know whether he stayed in Augusta after that."

Figure 5: Cross-lingual similarity scores calculated by XLM-R + Multiple or XLM-R + Single for two randomly chosen examples. The example sentences are shown in the sub-figure caption. Non-English-centric language pairs are highlighted with red color.

4.2 Monolingual Sentence Embeddings

SimCSE (Gao et al., 2021) exploits the ability of CL by using dropout as augmentation in unsupervised settings. For supervised sentence embeddings, SimCSE makes use of the NLI dataset (Bowman et al., 2015) to establish positive and hard negative examples. The success of SimCSE attracts researchers' attention to CL when dealing with sentence embeddings (Ni et al., 2022; Wang et al., 2022a; Su et al., 2022; Xie et al., 2022) since CL-based models provide more competitive results on downstream tasks than classical models such as S-BERT (Sentence-BERT) (Reimers and Gurevych, 2019). Su et al. (2022) further combine CL with prompt-like instructions while Liu et al. (2023) leverages ranking information into CL.

For now, few works exist considering multiple positives for monolingual sentence embeddings. Wu et al. (2022) initially expands a single positive instance into multiple positives for one anchor sentence through multiple augmentations.

4.3 Multilingual Sentence Embeddings

Training a universal sentence embedding model for all languages is a fundamental and important task. As the ability of efficient similarity calculation across languages, multilingual sentence embeddings have been applied to low-resource corpus filtering (Chaudhary et al., 2019), parallel corpus mining (Kvapilıková et al., 2021), and synthesized dataset filtering (Wu et al., 2023). Artetxe and Schwenk (2019) utilize the BiLSTM structure with a shared vocabulary for all languages. Recent studies prefer to adopt CL in multilingual settings with source-target translation pairs where the target sentence would be considered as the particular positive for the source sentence. For example, LaBSE (Feng et al., 2020) is a BERT-based multilingual sentence representation model trained on massive amounts of monolingual data and translation pairs covering over 100 languages. Mao and Nakagawa (2023) distill LaBSE for lightweight mutil-lingual sentence embedding models. Wang et al. (2022b) show that CL resembling SimCSE can also be applied to multilingual settings by using multilingual data. Besides, sentence-level CL is often combined with token-level information, for example, token-level reconstruction (Mao et al., 2022) and token-level alignment Li et al. (2023) to improve cross-lingual sentence embeddings. Although CL-based models have become common, leveraging multiple positives for learning multilingual sentence embeddings is still unexplored.

5 Conclusion

In this work, we propose MPCL, which improves multilingual sentence embeddings by utilizing a set of positive examples, specifically multilingual translations, for each anchor sentence. Through the incorporation of multiple positives, MPCL captures both linguistic diversity and transitive similarities, thereby enriching the embedding space. It updates pairwise similarity distributions into group-wise similarity ones by contrasting the anchor with multiple positives and employs a novel multi-positive loss function that simultaneously learns contrastive correlations and structural information among translations. By doing so, MPCL can improve performance in semantic similarity, retrieval, and classification tasks while exhibiting better robustness during training. Moreover, it is also observed in the experiments that MPCL shows a better transfer ability on unseen languages.

Limitations

Although we explore the effect of using multilingual translations as multiple positives, the experiments are still limited by the number of languages. Study on more low-resource languages could be taken into consideration. Besides, the XNLI data we used are machine-translated, with possible noises within. The composition of the training languages and how the training languages can affect the testing on other languages also remains to be explored. Additionally, as we study the impact of multiple multilingual positives for sentence embedding in this paper, the positives are assigned with equal importance. But there are various fine-grained strategies to weight positives such as ranking deserving exploited.

Ethics Statement

This paper attempts to improve existing sentence embedding approaches. All the data we used are open-sourced and contain no privacy-related ones. Our approaches are based on previously released codebases and checkpoints. We respect all work related to this work and expand ours on their well-established work. Our work does not introduce ethical biases but aims to make new, positive contributions to the multilingual computational languages community.

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*, 7:597–610.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Vishrav Chaudhary, Yuqing Tang, Francisco Guzmán, Holger Schwenk, and Philipp Koehn. 2019. Low-resource corpus filtering using multilingual sentence embeddings. *arXiv preprint arXiv:1906.08885*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan 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, Online. Association for Computational Linguistics.

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 1:539–546 vol. 1.

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, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Conference on Empirical Methods in Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Tejas Dhamecha, Rudra Murthy, Samarth Bharadwaj, Karthik Sankaranarayanan, and Pushpak Bhattacharyya. 2021. Role of Language Relatedness in Multilingual Fine-tuning of Language Models: A Case Study in Indo-Aryan Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8584–8595, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yangkai Du, Tengfei Ma, Lingfei Wu, Fangli Xu, Xuhong Zhang, Bo Long, and Shouling Ji. 2021. Constructing contrastive samples via summarization for text classification with limited annotations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1365–1376.

Xiangyu Duan, Hongfei Yu, Mingming Yin, Min Zhang, Weihua Luo, and Yue Zhang. 2019. Contrastive attention mechanism for abstractive sentence summarization. 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 3044–3053.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. 2021. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9588–9597.

Fangxiaoyu Feng, Yinfei Yang, Daniel Matthew Cer, N. Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. In *Annual Meeting of the Association for Computational Linguistics*.

Nicholas Frosst, Nicolas Papernot, and Geoffrey Hinton. 2019. Analyzing and improving representations with the soft nearest neighbor loss. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2012–2020. PMLR.

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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2019. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

Wataru Hirota, Yoshihiko Suhara, Behzad Golshan, and Wang-Chiew Tan. 2020. Emu: Enhancing multilingual sentence embeddings with semantic specialization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7935–7943.

David T Hoffmann, Nadine Behrmann, Juergen Gall, Thomas Brox, and Mehdi Noroozi. 2022. Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 897–905.

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 *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *Advances in Neural Information Processing Systems*, 33:21798–21809.

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.

Ivana Kvapilıková, Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Ondřej Bojar. 2021. Unsupervised multilingual sentence embeddings for parallel corpus mining. *arXiv preprint arXiv:2105.10419*.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021a. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. 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 2592–2607, Online. Association for Computational Linguistics.

Ziheng Li, Shaohan Huang, Zi qiang Zhang, Zhi-Hong Deng, Qiang Lou, Haizhen Huang, Jian Jiao, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. Dual-alignment pre-training for cross-lingual sentence embedding. In *Annual Meeting of the Association for Computational Linguistics*.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. RankCSE: Unsupervised sentence representations learning via learning to rank. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13785–13802, Toronto, Canada. Association for Computational Linguistics.

Zhuoyuan Mao, Chenhui Chu, and Sadao Kurohashi. 2022. Ems: Efficient and effective massively multilingual sentence representation learning.

Zhuoyuan Mao and Tetsuji Nakagawa. 2023. LEALLA: Learning lightweight language-agnostic sentence embeddings with knowledge distillation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1886–1894, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. 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 244–258.

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. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Hongjin Su, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, Tao Yu, et al. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

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

Yaushian Wang, Ashley Wu, and Graham Neubig. 2022b. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Qiyu Wu, Masaaki Nagata, and Yoshimasa Tsuruoka. 2023. WSPAlign: Word alignment pre-training via large-scale weakly supervised span prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11084–11099, Toronto, Canada. Association for Computational Linguistics.

Qiyu Wu, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, and Daxin Jiang. 2022. PCL: Peer-contrastive learning with diverse augmentations for unsupervised sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12052–12066, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742.

Yutao Xie, Qiyu Wu, Wei Chen, and Tengjiao Wang. 2022. Stable contrastive learning for self-supervised sentence embeddings with pseudo-siamese mutual learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:3046–3059.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94.

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, Vancouver, Canada. Association for Computational Linguistics.

A Appendix

A.1 Details of Figure 1

As shown in Figure 1, the similarity relationship in monolingual and multilingual are different. We calculate the similarity score of sentences by utilizing the released sample code and checkpoint of LaBSE [6]. Some monolingual examples we use can be found in Table 3 and multilingual examples in Table 4.

For monolingual examples, we randomly sample 100 sentences with entailments from the English XNLI dataset [7]. The premise is used as anchor and the corresponding entailment sentence serves as one of the positives. Other positives are generated by ChatGPT [8] and the prompt we used is "Give me several sentences that share similar meaning with the following one: ". For multilingual examples, we use the same 100 sentences and their corresponding translations from the XNLI dataset and calculate their similarity scores. We report the average similarity score of these sentences in Figure 1 (c) and (d).

A.2 Detailed Results of Main Experiments

We show detailed results of different tasks of in this section. BUCC's results is shown in Table 5. Some detailed results, especially languages involved in our training dataset are shown in Table 6. STS results can be found in Table 7. Full results of MTOP Domain Classification are shown in Table 8.

Model	fr	ru	zh	de	avg.
mBERT	-	-	-	-	56.7
+ Single	85.2	83.1	80.2	88.3	84.1
+ Multiple w/o hard negative	86.7	84.2	81.0	89.3	85.3
XLM-R	-	-	-	-	66.0
+ Single	94.2	95.1	93.1	95.2	94.5
+ Multiple w/o hard negative	94.9	96.0	95.2	96.4	95.7
+ Multiple w/ hard negative	94.5	95.0	96.6	95.5	95.4
mSimCSE$_{all}$	-	-	-	-	95.2
+Single	94.9	96.4	96.8	96.3	96.1
+Multiple	95.1	96.5	96.2	96.4	96.0
LaBSE	-	-	-	-	93.5
+Single	96.4	97.6	97.0	98.2	97.3
+Multiple	96.9	97.8	97.6	98.1	97.6

Table 5: Full results of BUCC task.

[6] https://huggingface.co/setu4993/LaBSE

[7] https://huggingface.co/datasets/xnli

[8] https://chat.openai.com/

Anchor Sentence	These organizations invest the time and effort to understand their processes and how those processes contribute to or hamper mission accomplishment.
Positive Instances	These organizations invest lots of time to understand how some processes can contribute to or hamper. In order to grasp the effects of those processes on the mission, time and effort are spent by these organizations. These associations dedicate to comprehending their procedure and assessing how they can either facilitate or impede their mission. Organizations spent much effort to understand the positive and negative impacts of their processes on the mission. To seize the influence of the processes towards the mission, associations sacrifice much time and effort.
Anchor Sentence	Thus, with respect to the litigation services Congress has funded, there is no alternative channel for expression of the advocacy Congress seeks to restrict.
Positive Instances	This is the only channel of expression of the advocacy that Congress seeks to restrict. Congress intends to curtail advocacy expression through this exclusive channel. This channel serves as the exclusive point of restriction for advocacy expression according to Congress. The funded litigation services represent the exclusive channel for the expression of the advocacy Congress seeks to restrict. With regard to Congress-funded litigation services, there are no alternative means for expressing the advocacy they intend to limit.
Anchor Sentence	My walkman broke so I'm upset now I just have to turn the stereo up real loud.
Positive Instances	I 'm upset that my walkman broke and now I have to turn the stereo up really loud. The broken walkman made me feel upset now and I'll turn loud the stereo. I'm feeling upset because my walkman is no longer functional and I'll turn loud the stereo. The broken walkman has left me feeling upset, and I have no other choice but to turn up the volume on the stereo. I'm feeling down because my walkman is broken thus I'll turn loud the stereo.

Table 3: Some examples that we use to calculate the similarity among mono-lingual multiple positive instances.

Anchor Sentence	These organizations invest the time and effort to understand their processes and how those processes contribute to or hamper mission accomplishment.
Positive Instances	Diese Organisationen investieren die Zeit und den Aufwand , um ihre Prozesse zu verstehen und wie diese Prozesse einen Beitrag zur Erfüllung der Aufgaben leisten oder behindern. Ces organisations investissent le temps et les efforts nécessaires pour comprendre leurs processus et la manière dont ces processus contribuent ou entravent la réalisation des missions. Estas organizaciones invierten el tiempo y el esfuerzo para comprender sus procesos y cómo esos procesos contribuyen o dificultan el logro de la misión. Эти организации вкладывают время и усилия для понимания своих процессов и того, каким образом эти процессы способствуют достижению целей миссии или препятствуют их достижению. 这些组织投入时间和努力来了解它们的进程以及这些进程如何有助于或妨碍特派团的成就.
Anchor Sentence	Thus, with respect to the litigation services Congress has funded, there is no alternative channel for expression of the advocacy Congress seeks to restrict.
Positive Instances	So gibt es in Bezug auf den Prozess der Rechtsstreitigkeiten, den der Kongress finanziert hat, keinen Alternativen Kanal für den Ausdruck des Advocacy-Kongresses zu beschränken. Por lo tanto, con respecto a los servicios judiciales que el congreso ha financiado, no existe ningún canal alternativo para la expresión del Congreso de promoción que pretende restringir. Ainsi, en ce qui concerne le congrès des services contentieux, il n'y a pas de voie alternative pour l' expression du congrès de plaidoyer. Таким образом , что касается деятельности конгресса по судебным услугам , то не существует какого-либо альтернативного канала для выражения мнений в рамках информационно-пропагандистского конгресса . 因此, 关于诉讼服务大会提供资金的问题, 没有任何其他渠道可以表达宣传大会试图加以限制的渠道.
Anchor Sentence	My walkman broke so I'm upset now I just have to turn the stereo up real loud.
Positive Instances	Mein Walkman ist kaputt , also bin ich sauer , jetzt muss ich nur noch die Stereoanlage ganz laut drehen . Mon Walkman S' est cassé alors je suis en colère maintenant je dois juste tourner la stéréo très fort Mi Walkman se rompió así que estoy molesto ahora solo tengo que girar el estéreo muy alto. Мой плеер сломался, так что я расстроен. Мне просто нужно включить стерео погромче. 我的随身听坏了所以我现在不高兴了我只能把立体声调大声.

Table 4: Some examples that we use to calculate the similarity among multi-lingual multiple positive instances.

Model	de	fr	es	ru	zh	avg$_{in}$	14 avg.	36 avg.
mBERT	-	-	-	-	-	-	-	-
+ Single	95.8	89.8	88.4	86.3	87.2	89.5	70.5	64.4
+ Multiple w/o hard negative	96.2	89.8	90.1	87.9	89.1	90.6	71.6	65.1
XLM-R	-	-	-	-	-	-	57.6	53.4
+ Single	98.6	95.0	97.6	93.6	95.6	96.1	91.3	89.6
+ Multiple w/o hard negative	99.0	95.2	97.8	94.1	96.5	96.5	92.0	90.4
+ Multiple w/ hard negative	98.9	95.6	97.8	94.4	96.2	96.6	93.5	91.8
mSimCSE$_{all}$ (Reproduced)	98.8	94.7	97.2	94.2	96.5	96.3	93.2	91.2
+ Single	98.8	95.6	98.1	93.8	96.4	96.5	93.3	91.2
+ Multiple	99.0	95.2	98.2	94.7	96.0	96.6	93.5	91.4
LaBSE (Reproduced)	99.0	96.0	96.3	95.3	96.1	96.5	95.3	95.0
+ Single	99.2	96.0	98.4	95.0	96.1	96.9	95.7	95.1
+ Multiple	99.4	96.4	98.3	95.2	96.4	97.1	96.0	95.4

Table 6: Detailed results of Tatoeba dataset. Since in the paper of LaBSE (Feng et al., 2020) and mSimCSE (Wang et al., 2022b), authors did not report the score for each language, we reproduce their scores on Tatoeba dataset through XTREME benchmark to have a better comparison. avg$_{in}$ stands for the languages included in our training dataset.

Model	STS17$_{in}$	STS17$_{ex}$	STS17 avg.	STS22$_{in}$	STS22$_{ex}$	STS22 avg.
mBERT	-	-	-	-	-	-
+ Single	63.7	51.5	57.0	54.3	52.2	53.4
+ Multiple	63.8	52.8	57.8	56.8	54.5	55.8
XLM-R	-	-	-	-	-	-
+ Single	73.8	68.9	71.1	61.2	57.8	59.8
+ Multiple w/o hard negative	74.8	71.9	73.2	63.2	58.9	61.4
+ Multiple w/ hard negative	81.1	77.1	78.9	65.1	62.5	64.0
mSimCSE$_{all}$ (Reproduced)	78.6	75.0	76.7	64.3	61.5	63.2
+ Single	79.1	75.1	76.9	63.6	59.8	62.0
+ Multiple	81.0	76.3	78.5	65.4	62.7	64.3
LaBSE	75.3	73.2	74.2	61.0	60.7	60.9
+ Single	76.1	74.6	75.3	61.0	57.8	59.7
+ Multiple	78.0	76.1	76.9	62.6	59.9	61.5

Table 7: Detailed results of STS tasks. We evaluate mSimCSE$_{all}$ through MTEB by ourselves. Task$_{in}$ stands for language pairs, that are inside our training set while Task$_{ex}$ stands for exclusive language pairs. More specifically, in STS17, there are five included pairs and six excluded pairs while in STS22, there are ten and seven, respectively (excluding *fr-pl*).

Model	de	en	es	fr	hi	th	avg.
mBERT	-	-	-	-	-	-	-
+ Single	74.6	78.0	75.1	69.3	59.8	16.9	62.3
+ Multiple	74.1	78.6	76.0	69.9	60.5	16.8	62.7
XLM-R	-	-	-	-	-	-	-
+ Single	84.5	85.4	85.0	81.7	79.2	82.1	83.0
+ Multiple w/o hard negative	86.3	86.2	86.8	82.3	83.1	82.0	84.5
+ Multiple w/ hard negative	88.0	88.5	88.8	86.2	85.9	83.8	86.8
mSimCSE$_{all}$ (Reproduced)	85.2	87.0	85.4	82.6	83.0	81.2	84.1
+ Single	86.6	87.7	87.4	83.8	83.3	82.1	85.2
+ Multiple	87.2	87.6	87.8	85.1	84.4	83.2	85.9
LaBSE	87.0	86.1	84.1	84.1	85.1	81.2	84.6
+ Single	87.2	88.0	86.2	84.5	85.7	82.0	85.6
+ Multiple	88.0	87.9	86.7	84.6	86.5	82.7	86.1

Table 8: Full results of MTOP domain classification. We report the accuracy metric on test set. Notice that *hi* and *th* are languages excluded from our training dataset.