English Contrastive Learning Can Learn Universal Cross-lingual Sentence Embeddings

Yau-Shian Wang **Ashley Wu** **Graham Neubig**
Carnegie Mellon University
king6101@gmail.com wangchew@andrew.cmu.edu
gneubig@cs.cmu.edu

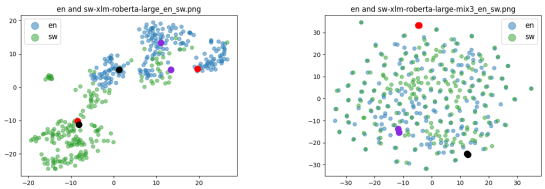
Abstract

Universal cross-lingual sentence embeddings map semantically similar cross-lingual sentences into a shared embedding space. Aligning cross-lingual sentence embeddings usually requires supervised cross-lingual parallel sentences. In this work, we propose mSimCSE, which extends SimCSE (Gao et al., 2021) to multilingual settings and reveal that contrastive learning on English data can surprisingly learn high-quality universal cross-lingual sentence embeddings without any parallel data. In unsupervised and weakly supervised settings, mSimCSE significantly improves previous sentence embedding methods on cross-lingual retrieval and multilingual STS tasks. The performance of unsupervised mSimCSE is comparable to fully supervised methods in retrieving low-resource languages and multilingual STS. The performance can be further enhanced when cross-lingual NLI data is available. [1]

1 Introduction

Universal cross-lingual sentence embeddings map the sentences from multiple languages into a shared embedding space, where semantically similar sentences across languages are close to each other. These embeddings have a wide spectrum of applications such as multi-lingual document retrieval (Artetxe and Schwenk, 2019a; Lin et al., 2020), multi-lingual question answering (Asai et al., 2021a,b; Kumar et al., 2022), unsupervised machine translation (Tran et al., 2020), and zero-shot transfer learning (Phang et al., 2020).

As shown in Figure 1 (a), without finetuning on downstream tasks, the embedding space of pre-trained multilingual language models such as m-BERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020) separate the embeddings of each language into different clusters. To align cross-lingual sentence embeddings, previous work (Artetxe and Schwenk, 2019a; Chidambaram et al., 2019; Feng et al., 2020) finetunes multilingual language models with billions of parallel data. However, it is non-trivial to obtain numerous parallel data for all languages. One potential direction to alleviate the need for parallel data is to enhance cross-lingual transfer of sentence embeddings.



(a) XLM-R without finetuning.

(b) XLM-R fintuned on English NLI data.

Figure 1: We visualize the sentence embeddings on XNLI corpus, where blue dots and green dots denote the sentences from English and Swahili respectively. Here, red dots, black dots, and purple dots denote the parallel sentences from different languages. In (a), the sentence embeddings from different languages are clearly separated into two clusters. In (b), after English NLI training, the embedding space becomes indistinguishable for different languages, and the parallel sentences are aligned to each other.

Pre-trained multilingual language models (Pires et al., 2019; Phang et al., 2020) have shown impressive performance on cross-lingual zero-shot transfer (Pires et al., 2019) that a model finetuned on a source language can generalize to target languages. This implies the representations finetuned on downstream tasks are universal across various languages. In this work, we explore various cross-lingual transfer settings on sentence retrieval tasks, especially in the setting of using English data only.

We propose multilingual-SimCSE (mSimCSE) which extends SimCSE (Gao et al., 2021), a famous sentence embedding method on English, to multilingual for cross-lingual transfer. SimCSE is

[1] Our code is publicly available at https://github.com/yaushian/mSimCSE.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9122 - 9133
December 7-11, 2022

a contrastive learning (Chopra et al., 2005; Hadsell et al., 2006; Chen et al., 2020a) method that pulls closer semantically similar sentences (i.e. positive sentence pairs) in embeddings space. As done in SimCSE, we obtain positive training pairs by either natural language inference (NLI) (Conneau et al., 2017; Reimers and Gurevych, 2019) supervision or unsupervised data augmentation using dropout. We also investigate model performance when a small amount of parallel data or cross-lingual NLI data are available.

In our experiments, as shown in Figure 1 (b), we are surprised to find that contrastive learning on pure English data seems to be able to align cross-language representations. Sentences that are semantically similar across languages are clearly closer together. Compared with previous unsupervised or weakly supervised methods, our unsupervised method significantly improves the performance on cross-lingual STS and sentence retrieval tasks. In retrieving low-resource languages and STS tasks, our method is even on par with fully supervised methods trained on billions of parallel data. Our results show that using contrastive learning to learn sentence relationships is more efficient than using massively parallel data for learning universal sentence embeddings. To the best of our knowledge, we are the first to demonstrate that using only English data can effectively learn universal sentence embeddings.

2 Related Work

2.1 Cross-lingual Zero-shot Transfer Learning

Due to the similarities between different languages, such as words, grammar, and semantics, multilingual models (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020; Wei et al., 2021; Chi et al., 2022) have been shown to generalize to unseen languages in a wide range of tasks. In machine translation, multilingual models exhibit the ability to perform zero-shot transfer in that they can translate on unseen language pairs (Zoph et al., 2016; Johnson et al., 2017). With the help of self-supervised learning, model can better acquire language-invariant representation, thus improving zero-shot machine translation (Siddhant et al., 2020; Liu et al., 2020).

To learn language-invariant representations cross-lingual tasks, previous work apply adversarial networks (Keung et al., 2019; Chen et al., 2019) or align representations via parallel corpus (Cao et al., 2020). Pires et al. (2019) revealed that mBERT is good at zero-shot cross-lingual transfer that fintuning on a specific monolingual task can generalize to other languages. More recently, X-Mixup (Yang et al., 2022) performs manifold mixup of source and target languages to learn general representations. Our method is the most relevant to the work which enhances cross-lingual transfer by using English downstream tasks to finetune multilingual language models (Phang et al., 2020) in that both methods leverage English NLI supervision. Our method differs from theirs in that we apply contrastive learning for representation learning and we only focus on sentence embeddings.

2.2 English Sentence Embeddings

Sentence embeddings aim to map sentences into a shared embedding space in which sentences with similar meanings can be close to each other. Previous work learns sentence embeddings by predicting the surrounding sentences of an input sentence, in either generative (Kiros et al., 2015) or discriminative (Logeswaran and Lee, 2018) manners. Recently, with the success of contrastive learning on learning visual representations (Chen et al., 2020b), more and more work explores contrastive learning on sentence embeddings. The training signal of contrastive learning can be obtained by data augmentation (Fang et al., 2020; Yan et al., 2021; Meng et al., 2021; Carlsson et al., 2021) or self-guided architecture (Kim et al., 2021; Carlsson et al., 2021). Among these, SimCSE (Gao et al., 2021) is the most famous one, which adopts either NLI supervision to define relevant sentences or unsupervised dropout as data augmentation, and improves state-of-the-art results. We choose to extend SimCSE to multilingual due to its impressive performance and simplicity.

2.3 Cross-lingual Sentence Embeddings

Universal cross-lingual sentence embeddings align semantically similar cross-lingual sentences into a shared embeddings space. To learn cross-lingual embeddings, previous work uses large amounts of parallel data to train neural networks to bring the embeddings of parallel sentences closer. LASER (Artetxe and Schwenk, 2019b) trains Bi-LSTM with parallel sentences of 93 languages to encourage consistency between the cross-lingual sentence embeddings. MUSE (Yang et al.,

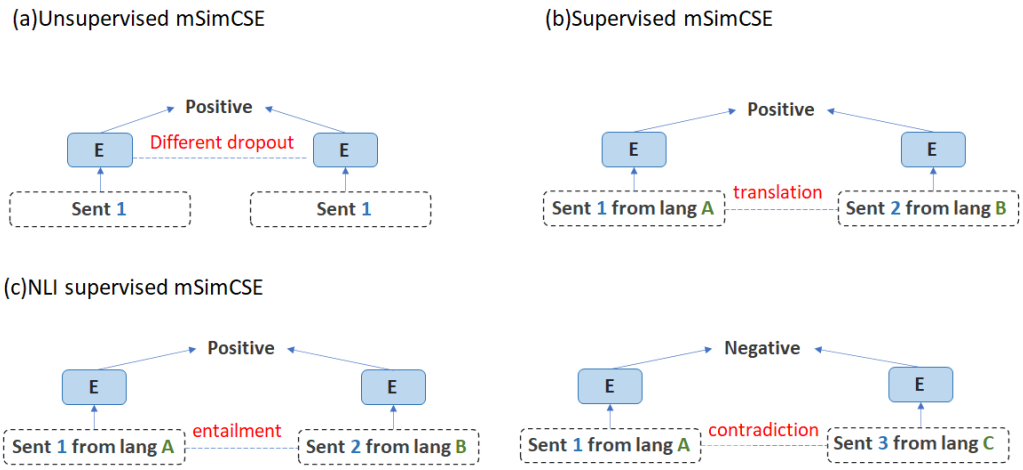


Figure 2: Overview of our method. In (a) unsupervised mSimCSE, sentence 1 is from English Wikipedia. It uses different dropout masks at each encoder inference as data augmentation. In (b) supervised SimCSE, we use parallel sentences as a positive training pair. In (c) NLI supervised mSimCSE, for the model only leverages English data, languages A, B, and C are all English. For mSimCSE that uses cross-lingual NLI supervision, languages A, B, and C are randomly sampled from a language pool. We use entailment and contradiction relationships between sentences to construct positive and hard negative training pairs.

2020) learns universal sentence embedding for 16 languages via translation based bridge tasks including multifeature question-answer prediction, translation ranking, and NLI. LaBSE (Feng et al., 2020) leverages dual-encoder framework to fine-tune mBERT on 6 billion parallel sentence pairs over 109 languages. Reimers and Gurevych (2020) extend the monolingual sentence embedding to multilingual by teacher-student training, in which a target language student model mimics source language embeddings from the teacher model. Sentence piece encoder (SP) (Wieting et al., 2021) simply learns sentence piece embeddings using parallel data and outperforms BERT-base models.

It is also possible to learn cross-lingual sentence embeddings in weakly supervised or even unsupervised manners. CRISS (Tran et al., 2020) uses the representations of mBART encoder as initial sentence embeddings, which are used to mine parallel texts from monolingual corpora. They iteratively extract bitexts and use the bitexts to update the model in a self-training loop. Notice that our method can be easily combined with CRISS by using mSimCSE as the initial model. Our method is closely related to DuEAM (Goswami et al., 2021), a dual-encoder model that leverages word mover distance and cross-lingual parallel NLI sentences to construct positive training pairs. The main differences are that we use contrastive learning, which is more effective than dual encoder, and that our method is more simple that only requires English data.

3 Method

3.1 SimCSE

SimCSE (Gao et al., 2021) apply batch contrastive learning (Chen et al., 2020b) to learn sentence embeddings on English data. Batch contrastive learning puts positive training pairs and negative training pairs into a same batch, increasing the difficulty of a contrastive task.

Specifically, a positive training pair means the sentences in the pair are semantically similar and their embeddings should be pulled closer, while a negative training pair means the sentences are semantically different. Given a batch of positive training pairs $B = \{(x_i, x_i^+)\}_{i=1}^N$, SimCSE calculates the batch constrastive loss for i-th pair as:

$$\mathcal{L}_i = -\log \frac{e^{sim(h_i,h_i^+)}}{\sum_{j=1}^N e^{sim(h_i,h_j^+)}}. \quad (1)$$

Here, N denotes the batch size, (x_i, x_i^+) denotes two semantically similar sentences, and $h_i = E(x_i)$ is the sentence embedding from encoder E. The key to using this loss is how to define the semantically similar pairs, which we elaborate on in the following section.

3.2 Multilingual SimCSE

In this section, we elaborate on how we extend SimCSE to multilingual and illustrate our proposed

mSimCSE in Figure 2. We explore four different multilingual training strategies, including the unsupervised strategy, the English NLI supervised strategy, the parallel NLI supervised strategy, and the fully supervised strategy. The difference between different strategies is how to define a positive training pair. Here, both unsupervised and English NLI supervised strategies can be recognized as an "unsupervised" setting for multilingual training because both of them only use English data and do not use any parallel data.

Unsupervised mSimCSE In unsupervised SimCSE, x_i and x_i^+ are the same sentence. As x_i and x_i^+ are encoded by the same encoder but with different dropout, the dropout can be viewed as a light-weight data augmentation method. We use the wikipedia data from the original SimCSE repository to train our model.

English NLI supervised mSimCSE In the English NLI supervised strategy, we use English natural language inference (NLI) (Conneau et al., 2017; Reimers and Gurevych, 2019) datasets to construct positive and hard negative training pairs. Specifically, if two sentences are labeled as "entailment" relationship, they are viewed as a positive pair (x_i, x_i^+). For each x_i, we also include a hard negative example x_i^- in the same training batch, where x_i and x_i^- are labeled as "contradiction" relationship.

Cross-lingual NLI supervised mSimCSE The English NLI supervision can be easily extended to the multilingual strategy by constructing a positive training pair from different languages. We use XNLI (Conneau et al., 2018), which translates English NLI to multiple languages. Similar to the English NLI strategy, the cross-lingual sentence pairs with "entailment" and "contradiction" relationship are viewed as positive and negative pairs respectively, but as shown in Figure 2, the language of x_i, x_i^+ and x_i^- are randomly sampled. They can come from either different languages or the same language.

Supervised mSimCSE In supervised mSimCSE, we simply define a positive training pair as the parallel sentences from different languages. This strategy is the same as previous supervised methods, but we use relatively few parallel sentences. Note that different strategies can be easily combined by mixing training pairs from different strategies.

4 Experiments

4.1 Experimental Setup

Training Details We adapt SimCSE codebase [2] to a multi-lingual setting. We keep all other hyperparameters same as the original SimCSE, and fix learning rate to be $1e-5$, training epoch to be 1, and batch size to be 128 for all experiments. We use our method to finetune XLM-Roberta-large (XLM-R) (Conneau et al., 2020). We examine the performance of different hyperparameters in Appendix A

Training Data for Different mSimCSE Strategies In unsupervised mSimCSE and English NLI supervised mSimCSE, we use the pre-processed English Wikipedia and English NLI training tuples downloaded from the SimCSE codebase respectively. In all the tables in this paper, the subscripts of $mSimCSE$ denote the languages that we use to train our model. In cross-lingual NLI supervision, $mSimCSE_{en,fr}$ denotes we use English and translated French NLI data to train our model and $mSimCSE_{all}$ means that we use all the languages in XNLI (Conneau et al., 2018) dataset.

In supervised finetuning, $mSimCSE_{sw}$ denotes that we use the translation pairs of English and Swahili. For each language, we randomly select 100k parallel sentences from ParaCrawl project [3] via the OPUS corpus collection [4]. In supervised finetuning, "$mSimCSE_{sw}$+NLI" denotes that we mix English NLI sentence pairs with English-Swahili translation pairs. Note that because parallel sentences don't have hard negative sentences, to mix them with NLI data, we also remove hard negative sentences of NLI in "$mSimCSE_{sw}$+NLI".

4.2 Baselines

First, we compare our method to unsupervised pretrained language models including XLM-R and M-BERT without finetuning to show that unsupervised contrastive learning can learn more generalized cross-lingual sentence embeddings. In some tasks, we also compare our method with more competitive language models in Xtreme (Hu et al., 2020) benchmark, such as XLM-E (Chi et al., 2022), HICTL (Wei et al., 2021) and INFOXLM (Chi et al., 2021). CRISS (Tran et al., 2020) is an unsupervised sentence retrieval method, which mines

[2] https://github.com/princeton-nlp/SimCSE

[3] http://paracrawl.eu

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

Models	BUCC	Tatoeba-14	Tatoeba-36
Unsupervised			
XLM-R	66.0	57.6	53.4
INFOXLM	-	77.8	67.3
DuEAM	77.2	-	-
XLM-E	-	72.3	62.3
HiCTL	68.4	-	59.7
$mSimCSE_{en}$	87.5	82.0	78.0
English NLI supervised			
(Phang et al., 2020)	71.9	-	81.2
$mSimCSE_{en}$	**93.6**	**89.9**	**87.7**
Cross-lingual NLI supervised			
$mSimCSE_{en,fr}$	94.2	90.8	88.8
$mSimCSE_{en,fr,sw}$	94.3	93.3	90.3
$mSimCSE_{all}$	**95.2**	93.2	91.4
DuEAM	81.7	-	-
Fully Supervised			
LASER	92.9	95.3	84.4
LaBSE	93.5	95.3	95.0
$mSimCSE_{sw}$	86.8	87.7	86.3
$mSimCSE_{fr}$	87.1	87.9	85.9
$mSimCSE_{sw,fr}$	88.8	90.2	88.3
$mSimCSE_{sw,fr}$+NLI	93.6	91.9	90.0

Table 1: Results of sentence retrieval task on Xtreme benchmark. We report F1-scores for BUCC and accuracy for Tatoeba.

parallel sentences from multiple monolingual corpora using self-training.

Our main baselines are other methods which also leverage NLI supervision. Phang et al. (2020) finetune multilingual language models on various English tasks, including English NLI task. DuEAM (Goswami et al., 2021) also leverages multilingual NLI supervision to learn universal sentence embeddings.

We also compare our method with fully supervised methods that leverages parallel sentences, including LASER (Artetxe and Schwenk, 2019b), LaBSE (Feng et al., 2020), and SP (Wieting et al., 2021). Note that among all the methods, $mSimCSE_{en}$ is the only method that only uses English data.

4.3 Sentence Retrieval

Following the previous work of cross-lingual sentence embedding learning (Goswami et al., 2021), we evaluate our model on multi-lingual sentence retrieval, including Tatoeba (Artetxe and Schwenk, 2019b) and BUCC (Zweigenbaum et al., 2018). Tatoeba requires models to match parallel sentences from source and target language sentence pools. BUCC is a bitext mining task, in which a model needs to rank all the possible sentence pairs, and predicts sentence pairs whose scores are above an optimized threshold.

In Table 1, we follow the setting in Xtreme benchmark to evaluate model performance on sentence retrieval task. In unsupervised setting, mSimCSE trained on English Wikipedia improves the performance by a large margin. This implies that contrastive training can effectively pull closer cross-lingual semantically similar sentences.

With English NLI supervision, it significantly improves the performance against unsupervised methods and DuEAM. It even beats fully-suprvised methods that leverages parallel sentences on BUCC task. This implies that with an objective that learns more difficult semantic relationship between sentences, model can learn better universal cross-lingual sentence embeddings.

By comparing cross-lingual NLI supervised mSimCSE with Enlgish NLI supervised mSimCSE, we observe that the model performance can be further improved using translated NLI pairs from other languages. In general, including more languages can improve the performance. Comparing with DuEAM which also leverages parallel NLI supervision, contrastive learning can learn universal sentence embedding more effectively. In the BUCC dataset, the performance of $mSimCSE_{all}$ is better than fully supervised methods, such as LaBSE, which is trained on 6 billion parallel data. Note that $mSimCSE_{all}$ uses far less parallel data than LaBSE, which demonstrates the effectiveness of our method.

In the fully supervised setting, comparing

Models	hi	fr	de	af	te	tl	ga	ka	am	sw
Unsupervised										
CRISS	92.2	92.7	98.0	-	-	-	-	-	-	-
DuEAM	83.5	-	93.4	79.9	78.6	56.8	35.0	70.7	46.4	-
$mSimCSE_{en}$	86.9	87.2	94.1	76.0	78.8	49.7	39.2	75.2	48.8	29.4
English NLI supervised										
$mSimCSE_{en}$	**94.4**	**93.9**	**98.6**	**85.6**	**92.9**	**70.0**	**54.8**	**89.2**	**79.5**	**42.1**
Cross-lingual NLI supervised										
DuEAM	92.9	-	96.0	84.8	90.6	60.6	42.0	76.4	56.0	-
$mSimCSE_{en,fr}$	95.1	94.4	98.8	88.9	94.2	73.4	59.4	91.3	79.5	44.5
$mSimCSE_{en,fr,sw}$	95.7	94.2	98.4	87.9	94.4	75.6	62.1	90.5	82.7	**75.5**
$mSimCSE_{all}$	**96.2**	94.8	98.8	**90.6**	**96.2**	**80.9**	**65.1**	**92.4**	**82.4**	67.8
Fully supervised										
LASER	94.7	95.7	99.0	89.4	79.7	-	5.2	35.9	42.0	42.4
$mSimCSE_{sw}$	94.3	91.6	97.6	85.2	88.5	76.3	60.8	85.5	65.2	47.6
$mSimCSE_{fr}$	94.1	92.6	97.3	84.6	89.3	70.8	54.6	86.3	63.4	43.6
$mSimCSE_{sw,fr}$	95.1	93.8	97.8	86.1	91.2	75.8	59.6	88.9	74.4	51.5
$mSimCSE_{sw,fr}$+NLI	95.8	94.7	98.6	89.8	95.7	77.8	63.9	91.7	81.0	57.1

Table 2: Accuracy of Tatoeba multilingual retrieval task.

$mSimCSE_{sw,fr}$ with and without NLI supervision, we find that if the translation pairs are rare, adding English NLI supervision can significantly improve the performance. Also, compared with the mSimCSE that only uses English NLI supervision, adding a few extra parallel data can slightly improve the performance.

Following DuEAM, in Table 2, we select some high-resource languages including Hindi (hin), French (fra), German (deu), Afrikaans (afr) and Swahili (swh), and low-resource languages including Telugu (tel), Tagalog (tgl), Irish (gle), Georgian (kat), and Amharic (amh) from Tatoeba dataset to further analyze the model performance. In high-resource languages, fully-supervised achieves better performance because large amounts of parallel sentence pairs are available in these languages. On the other hand, in low-resource languages, due to the lack of training pairs, supervised method can not generalize well on these languages while mSimCSE can generalize better.

Finally, we evaluate whether including cross-lingual NLI supervision in a target language can improve the performance. In Table 2, compared with using only English NLI supervision, $mSimCSE_{en,fr,sw}$ in the cross-lingual NLI setting which includes Swahili in training significantly improves the performance in Swahili. Its performance gain is greater than fully supervised "$mSimCSE_{sw,fr}$+NLI", which leverages parallel sentences.

Models	ar-ar	ar-en	es-es	es-en	tr-en
Unsupervised					
XLM-R	53.5	26.2	68.1	10.7	10.5
mBERT	55.2	28.3	68.0	23.6	17.3
$mSimCSE_{en}$	72.3	48.4	83.7	57.6	53.4
English NLI supervised					
$mSimCSE_{en}$	**81.6**	71.5	**87.5**	**79.6**	71.1
Cross-lingual NLI supervised					
DuEAM	69.7	54.3	78.6	56.5	58.4
$mSimCSE_{all}$	79.4	72.1	85.3	77.8	74.2
Fully Supervised					
LASER	79.7	-	57.9	-	72.0
LaBSE	80.8	-	65.5	-	72.0
SP	76.7	78.4	85.6	77.9	79.5
$mSimCSE_{sw,fr}$+NLI	77.7	72.4	86.3	79.7	72.5

Table 3: Spearman rank correlation (ρ) results for SemEval 2017 STS shared task. The results of supervised baselines and DuEAM are taken from (Goswami et al., 2021).

4.4 Cross-lingual STS

Cross-lingual STS (Cer et al., 2017) evaluates whether a model predicted semantic similarity between two sentences are correlated to human judgement. The two sentences can come from either the same language or different languages. Given a sentence pair, we compute the cosine similarity of sentence embeddings as a model prediction.

The results of multi-lingual STS benchmark are shown in Table 3. For unsupervised XLM-R and mBERT without finetuning, we try several pooling methods and find that averaging over the first and the last layers yields the best results on STS. The poor results of pre-trained language models mean that the sentence embeddings of pre-trained language models do not capture the semantics in the cosine similarity space well. With unsupervised mSimCSE pre-training, it enhances the semantics of monolingual STS tasks, i.e. "ar-ar" and "es-es".

For the tasks that requires cross-lingual alignment, including "ar-en", "en-es", and "tr-en", the gap between unsupervised baselines and English NLI supervised baselines is still large.

Comparing with the methods that utilize either parallel NLI supervision or supervised parallel data, the English NLI training achieves the best results on ar-ar, es-es, and es-en pairs. This implies that mSimCSE can effectively transfer semantic knowledge learned from English NLI supervision to other languages. We find that using parallel data can only improve the results of bilingual pairs, and reduces the performance of monolingual pairs.

4.5 Unsupervised Classification

We conduct unsupervised classification to evaluate whether the model can cluster semantically similar documents together on the languages other than English. We use Tnews dataset in CLUE benchmark [5] to evaluate the performance of our model. Tnews is a Chinese news classification dataset, which contains 15 news categories. We first conduct k-means clustering on sentence embedding, in which cluster number k is set to be the same as the number of news categories. Then, we evaluate the mean accuracy of each cluster that measures what percentages of documents in each cluster are from the same human-labeled category.

Compared with unsupervised pre-trained language models, mSimCSE significantly improves the purity scores. This is expected because without fine-tuning, the embeddings from pre-trained language models cannot capture relative distances between instances. Similar to the observation in the previous section, English NLI supervision can greatly enhance the performance, closing the gap between the fully supervised fine-tuned BERT.

4.6 Zero-shot Cross-lingual Transfer of Sentence Classification

To evaluate the cross-lingual zero-shot transfer of pre-trained sentence embedding, we evaluate our model on PAXS-X (Yang et al., 2019) sentence classification task. PAXS-X requires a model to determine whether two sentences are paraphrases. In Table 5, compared with XLM-R and XLM-E (Chi et al., 2022), mSimCSE without using NLI data improves the performance, which demonstrates that mSimCSE is an effective approach for zero-shot cross-lingual transfer. In this task, using English NLI supervision does not improve performance.

[5] https://github.com/CLUEbenchmark/CLUE

Models	Purity
Unsupervised	
Random	6.7
mBERT	15.2
XLM-R	13.7
$mSimCSE_{en}$	30.3
English NLI supervision	
$mSimCSE_{en}$	**40.3**
Cross-lingual NLI supervision	
$mSimCSE_{all}$	41.6
Supervised Classification Model	
BERT	56.6

Table 4: Accuracy of unsupervised clustering on Tnews classification dataset. In supervised fintuning, the model is finetuned on classification training set.

Models	Accuracy
Unsupervised	
mBERT	81.9
XLM-R	86.4
XLM-E	87.1
$mSimCSE_{en}$	**88.1**
English NLI supervised	
(Phang et al., 2020)	87.9
$mSimCSE_{en}$	88.2

Table 5: Accuracy on PAWS-X dataset.

5 Analysis

5.1 The Effect of Parallel Sentences Number

Parallel data	BUCC	Tatoeba14	Tatoeba36
0	91.4	90.4	88.0
10k	92.6	90.6	88.5
100k	93.5	90.8	88.6
1M	94.4	90.6	88.5
5M	94.5	90.7	88.2

Table 6: The effect of parallel English-French sentences number.

As parallel data is easy to obtain for most languages (Artetxe et al., 2020), we investigate the effect of the number of parallel sentences. We mix parallel English-French sentences with the English NLI data and gradually increase the number of parallel sentences. Here, 0 parallel sentence means we only use English NLI data without using hard negative examples, so the results are different from the results in Table 1.

The BUCC dataset has only four high-resource languages, of which French is one of them. With more parallel data, the consistent improvement on BUCC dataset implies that using more English-

French translation pairs can improve the performance on the English-French mining task, thus improving the results of BUCC. On the other hand, the Tatoeba dataset includes much more languages, which evaluates a model's generalization. We observe that using more parallel data does not influence the performance on Tatoeba, which implies that using translation data on a single language pair does not generalize well to other languages. The results suggest that using large amounts of parallel data may not be the most efficient way to learn universal sentence embeddings while learning sentence relationships is a more promising direction.

5.2 Can Contrastive Learning Removes Language Identity?

Models	en,de,fr,hi	en,tr,ar,bg
XLM-R	99.2	99.8
$mSimCSE_{en}$	91.1	95.3

Table 7: Accuracy of language classifier.

Masked language modeling requires a model to capture the language identity to predict correct tokens for a specific language. On the other hand, English contrastive loss only learns the relationship between sentences, which does not seem to require language identity. We speculate that the contrastive loss can thus remove the language identity in sentence embeddings and enhance the general shared cross-lingual semantics.

To verify this, in Table 7, we train two language classifiers on en,de,fr,hi and en,tr,ar,bg respectively. The language classifier needs to predict the correct language of the input sentence embeddings. We use the sentences from XNLI as our training and testing data. With contrastive learning, the accuracy of language classifier decreases, which implies the embeddings are more language-invariant, which more or less verifies our assumption. However, the accuracy is still very high because the language classifier can still predict the language of a text by language-specific features such as grammar and characters.

6 Discussion

Our experimental results demonstrate that in both unsupervised and English NLI supervised settings, using English data alone can surprisingly align cross-lingual sentence embeddings. By comparing unsupervised results with NLI supervised results, we observe that learning more meaningful sentence relationships can further enhance the alignment. In our analysis, we find that infinitely increasing parallel training data is not the most efficient manner to learn universal sentence embeddings; instead, our results suggest that designing a more challenging contrastive task or more effective sentence embedding learning method on English data may be a more efficient direction. Also, contrastive learning may be a promising direction for improving the zero-shot transfer of pre-trained multilingual language models.

We attribute the alignment to language-invariant contrastive training. Because multilingual language models have shown good performance on zero-shot transfer, we speculate that multilingual language models encode texts into two disentangled embeddings, a language-specific embedding and a general language agnostic embedding. Because English contrastive task doesn't require mlms to capture language identity, it only pulls closer language-agnostic sentence embeddings while weakening language-specific embedding. This property can be verified in Figure 1 and Table 7. However, it still requires more investigation to fully understand why contrastive learning on English data can achieve cross-lingual transfer.

7 Conclusion

In this work, we demonstrate that using only English data can effectively learn universal sentence embeddings. We propose four different strategies to extend SimCSE to multilingual, including unsupervised, English NLI supervised, cross-lingual NLI supervised, and supervised strategies. We surprisingly find that the English NLI supervised strategy can achieve performance on par with previous supervised methods that leverage large amounts of parallel data. Our work provides a new perspective on learning universal sentence embeddings and cross-lingual transfer that using contrastive learning to learn sentence semantic relationships on monolingual corpora may be a promising direction.

8 Limitations

In the previous sections, we attribute why models trained on English data can learn cross-lingual sentence embeddings to language-invariant contrastive task. We speculate that multilingual language models have already implicitly learned such universal

representations, but they also learn some language-specific representations. Contrastive learning enhances the language-invariant representations, diminishing the language-specific representations without distorting the semantics embedded in the representations. However, this speculation still requires more evidence to support it. Also, it is important to understand to which the zero-shot transfer happens, such as which languages are easier to transfer, and what is the properties of these languages. For the properties of these languages, by observing the experimental results in Table 2, we have two speculations, one is their similarity between English, and another one is the number of the monolingual pre-training data for these languages, but again these speculations also require more analysis to verify. By understanding the reason for this phenomenon, it is possible to achieve better zero-shot transfer and learn more "universal" sentence embeddings.

References

Mikel Artetxe, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre. 2020. A call for more rigor in unsupervised cross-lingual learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7375–7388, Online. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019a. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hanna Hajishirzi. 2021b. One question answering model for many languages with cross-lingual dense passage retrieval. In *Advances in Neural Information Processing Systems*, volume 34, pages 7547–7560. Curran Associates, Inc.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A simple framework for contrastive learning of visual representations.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. 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.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: Cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6170–6182, Dublin, Ireland. Association for Computational Linguistics.

Muthu Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259, Florence, Italy. Association for Computational Linguistics.

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application

to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 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, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding.

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

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John Philip McCrae. 2021. Cross-lingual sentence embedding using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9099–9113.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR '06, page 1735–1742, USA. IEEE Computer Society.

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 generalization. *CoRR*, abs/2003.11080.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Phillip Keung, Yichao Lu, and Vikas Bhardwaj. 2019. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER. 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 1355–1360, Hong Kong, China. Association for Computational Linguistics.

Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2021. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540, Online. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Gokul Karthik Kumar, Abhishek Singh Gehlot, Sahal Shaji Mullappilly, and Karthik Nandakumar. 2022. Mucot: Multilingual contrastive training for question-answering in low-resource languages. *arXiv preprint arXiv:2204.05814*.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: Bert and beyond.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Yu Meng, Chenyan Xiong, Payal Bajaj, saurabh tiwary, Paul Bennett, Jiawei Han, and XIA SONG. 2021. Coco-lm: Correcting and contrasting text sequences for language model pretraining. In *Advances in Neural Information Processing Systems*, volume 34, pages 23102–23114. Curran Associates, Inc.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575, Suzhou, China. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2827–2835, Online. Association for Computational Linguistics.

Chau Tran, Yuqing Tang, Xian Li, and Jiatao Gu. 2020. Cross-lingual retrieval for iterative self-supervised training. In *Advances in Neural Information Processing Systems*, volume 33, pages 2207–2219. Curran Associates, Inc.

Xiangpeng Wei, Rongxiang Weng, Yue Hu, Luxi Xing, Heng Yu, and Weihua Luo. 2021. On learning universal representations across languages. In *International Conference on Learning Representations*.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Paraphrastic representations at scale. *arXiv preprint arXiv:2104.15114*.

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

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *International Conference on Learning Representations*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 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, Online. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of EMNLP 2019*, pages 3685–3690.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of 11th Workshop on Building and Using Comparable Corpora*, pages 39–42.

epoch	bs	lr	BUCC	Tatoeba-14	Tatoeba-36
1	128	1e-5	93.6	89.9	87.7
2	128	1e-5	94.4	90.0	87.8
3	128	1e-5	93.5	90.0	87.4
1	256	1e-5	93.3	90.0	87.9
1	128	2e-5	93.7	90.0	87.4

Table 8: Performance of different Hyperparameters. Epoch denotes training epoch number, bs denotes the batch size, and lr denotes learning rate.

A Hyperparameters

In Table 8, we show how different hyperparameters influence the model performance. We choose $mSimCSE_{en}$ trained on English NLI data as the model under examination. We find that the performance of different hyperparameters is very close, which implies that our method is stable and not sensitive to hyperparameters. Increasing the number of training epochs to 2 can improve the performance on BUCC.