Semantic Alignment with Calibrated Similarity for Multilingual Sentence Embedding

Jiyeon Ham[1] **Eun-Sol Kim**[1,2]
[1]Kakao Brain
[2]Hanyang University
jiyeon.ham@kakaobrain.com, eunsolkim@hanyang.ac.kr

Abstract

Measuring the similarity score between a pair of sentences in different languages is the essential requisite for multilingual sentence embedding methods. Predicting the similarity score consists of two sub-tasks, which are monolingual similarity evaluation and multilingual sentence retrieval. However, conventional methods have mainly tackled only one of the sub-tasks and therefore showed biased performances. In this paper, we suggest a novel and strong method for multilingual sentence embedding, which shows performance improvement on both sub-tasks, consequently resulting in robust predictions of multilingual similarity scores. The suggested method consists of two parts: to learn semantic similarity of sentences in the pivot language and then to extend the learned semantic structure to different languages. To align semantic structures across different languages, we introduce a teacher-student network. The teacher network distills the knowledge of the pivot language to different languages of the student network. During the distillation, the parameters of the teacher network are updated with the slow-moving average. Together with the distillation and the parameter updating, the semantic structure of the student network can be directly aligned across different languages while preserving the ability to measure the semantic similarity. Thus, the multilingual training method drives performance improvement on multilingual similarity evaluation. The suggested model achieves the state-of-the-art performance on extended STS 2017 multilingual similarity evaluation as well as two sub-tasks, which are extended STS 2017 monolingual similarity evaluation and Tatoeba multilingual retrieval in 14 languages.

1 Introduction

Representing semantics of sentences as embedding vectors on a vector space is crucial for various natural language processing (NLP) tasks. The fundamental inductive bias of the sentence embedding methods is to place sentences having similar semantics close to each other in the vector space, which is advantageous to sentence-based tasks such as clustering and semantic retrieval. Building upon well-known pre-trained English language models BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), Reimers and Gurevych (2019) introduced a fine-tuning method to learn the semantic similarity between two English sentences using siamese networks. However, it has focused only on monolingual settings.

Multilingual sentence embedding models should be able to measure the semantic similarity between a pair of sentences not only in the same language but also in different languages. There are two conditions to satisfy when measuring the similarity between a pair of sentences in different languages. Firstly, monolingual sentences need to be closely placed as similar as they are. Secondly, a translation pair, which have the same meaning in different languages, should be placed in close proximity to each other. For example, in figure 1, the similarity score between a) and b) is 0.4, and a Korean translation of b) is c). If two English sentences are positioned to express a similarity score of 0.4, and the translation pair are placed very close, we can measure the similarity score between a) and c) as around 0.4. Therefore, satisfying the two conditions enables the model to calculate similarity scores of multilingual sentences. The first condition can be evaluated by the monolingual sentence similarity evaluation task, and the second condition can be assessed by a multilingual sentence retrieval task.

Even though several methods have been recently proposed for multilingual sentence embedding, they fail to achieve high performance on the multilingual sentence similarity evaluation because they have not succeeded in reaching both conditions. LASER (Artetxe and Schwenk, 2019) and

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1781–1791
November 7–11, 2021.

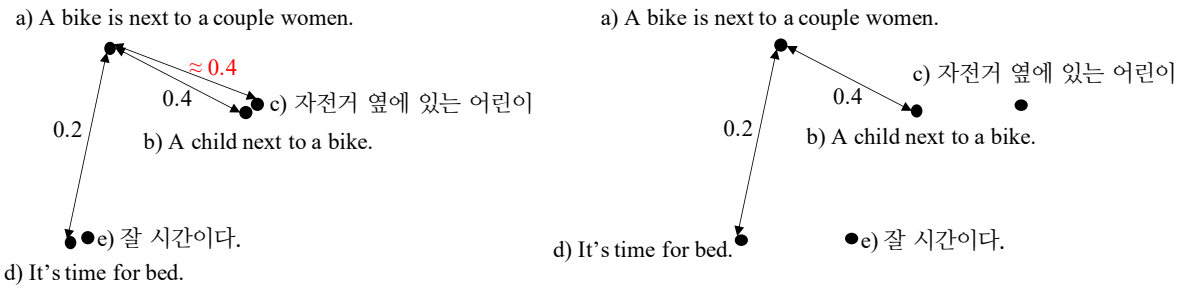

Figure 1: Multilingual sentence embedding methods are compared with five examples (a-e). Two kinds of the dataset have been widely used, which are semantic similarity scores for English sentences (0.4 for a and b, 0.2 for a and d) and translation dataset (b-c, d-e). The key idea of the suggested method is to learn the semantic structure of pivot language (English) with the first dataset and then extend the structure to different languages by aligning embeddings of translation pairs directly (left). As the left figure shows, the suggested method can correctly measure similarities between different languages (0.4 for a and c) because embeddings of b and c are closely aligned and the semantic relationship between a and b is trained. On the other hand, LASER (Artetxe and Schwenk, 2019) and LaBSE (Feng et al., 2020) suggested contrastive objectives to keep alignment between translation pairs (b-c, d-e), thus shows inferior performance on similarity evaluation tasks compared to their retrieval performance. Recent work Reimers and Gurevych (2020) also suggested a method to align semantic structures across different languages. However, their alignment method is indirect (right), thus shows inferior performance on the retrieval.

LaBSE (Feng et al., 2020), which are trained with contrastive objectives using translation pairs, show inferior performance on the similarity evaluation tasks compared to their retrieval accuracy. In Figure 1, they focus on aligning b) and c), but not a) and b). A model proposed by Reimers and Gurevych (2020) is successful at measuring the similarity of monolingual sentences but shows a performance drop in the retrieval task because the model less directly aligns sentence vectors across the different languages. In other words, as shown in Figure 1, the model Reimers and Gurevych (2020) embeds a) and b) to correctly present the similarity score 0.4, but is unsuccessful to represent the relationship between b) and c).

In this paper, we introduce a powerful multilingual sentence embedding model which is able to measure similarity between multilingual sentences. The main idea of the suggested model is to align semantics across different languages after training on a monolingual semantic similarity dataset. First of all, the model is trained to measure the similarity between two sentences in the pivot language using a semantic similarity dataset. To extend the monolingual model to multilingual settings, we suggest a teacher-student network architecture. The student network, which is the final model of our approach, should align sentence vectors across multiple languages while keeping the learned semantic structure. The teacher network, which captures the semantic similarity in the pivot language, distills the semantics of the pivot language to other languages of the student network using translation pairs. In the meantime, the teacher network is not fixed but slowly adapted to the student network. The distillation and the adaptation together enable the alignment between the pivot language and the other language of the student network. Finally, the student network produces the multilingual sentence embedding that can measure the calibrated similarity which can not only determine whether a translation pair are close but also quantify the semantic similarity. We demonstrate that the suggested method achieves the state-of-the-art performance on multilingual sentence similarity evaluation of extended STS 2017 (Reimers and Gurevych, 2020) as well as STS 2017 monolingual similarity evaluation and Tatoeba-14 languages (Artetxe and Schwenk, 2019) multilingual sentence retrieval.

2 Related Work

2.1 Monolingual Sentence Embedding Models

SBERT (Reimers and Gurevych, 2019) is a state-of-the-art sentence embedding model on STS benchmark dataset (Cer et al., 2017) in English. As the usefulness of natrual language inference (NLI) data (Bowman et al., 2015; Williams et al., 2018) known from InferSent (Conneau et al., 2017), they trained the BERT model on NLI data and then trained on STS benchmark data. The whole training process was done using the siamese network

structure.

While SBERT proposed training strategy, Augmented SBERT (Thakur et al., 2020) introduced a data augmentation method to achieve high scores on argument similarity, semantic textual similarity, duplicate question detection, and news paraphrase identification. Cross-encoders, which encode input sentences jointly, often perform better than bi-encoders, which encode input sentences separately. For the tasks that cross-encoders perform better than bi-encoders, they labeled additional training pairs using cross-encoders to train bi-encoders as sentence embedding models.

2.2 Multilingual Sentence Embedding Models

LASER (Artetxe and Schwenk, 2019) trained LSTM-based encoder-decoder model using translation task. They used only the encoder to generate a sentence vector. LaBSE (Feng et al., 2020) is a multilingual BERT-based model trained on translation pairs by additive margin softmax. They regarded translation pairs to the positive samples and the other in-batch samples to the negative. Multilingual universal sentence encoder (m-USE) (Yang et al., 2019) trained a single shared encoder on multiple tasks such as NLI, QA, translation ranking. Reimers and Gurevych (2020) proposed to train a multilingual student model, which is started from the multilingual pre-trained MLM model named XLM-R (Conneau et al., 2020), by the distillation of the English SBERT teacher model. Because the teacher model can assess the similarity of sentences, the student model learns to compare the similarity.

2.3 Representation Learning

While BERT succeeded in NLP by self-supervised learning using a masked language model, contrastive loss methods are in the limelight of self-supervised learning in vision (Chen et al., 2020; He et al., 2020; Oord et al., 2018; Tian et al., 2019, 2020). These methods close the distance between the representation of different augmented views from the same image and broaden the distance between the representation of augmented views from the different images. However, the performance of these methods often relies on the size or quality of negative samples.

Bootstrap your own latent (BYOL) method (Grill et al., 2020) is a self-supervised image representation learning method using only positive samples. Using two neural networks, referred to as online and target, they train the online network to predict the target network's representation. At the same time, they receive different augmented views of the same image. The gradient only updates the online network, and the target network slowly updates from the online network's parameters.

Furthermore, as MoCo (He et al., 2020) and BYOL update parameters of sub-module with different rates, Zhang and Khoreva (2019) updates parameters of the generator and the discriminator with different learning rates in GAN (Goodfellow et al., 2014).

3 Tasks

We conduct three tasks to evaluate the semantic alignment of the multilingual sentence embedding model: multilingual similarity evaluation, monolingual similarity evaluation, and multilingual sentence retrieval tasks.

3.1 Multilingual and Monolingual Similarity Evaluation Tasks

The task measures the similarity between not only the sentences with similar meanings but also the sentences with dissimilar meanings.

For the multilingual and the monolingual similarity evaluation tasks, we adopt extended STS 2017 dataset (Reimers and Gurevych, 2020). They provide labels of the similarity between two sentences from 0 (no meaning overlap) to 5 (equivalent meaning), which is annotated by humans. A sentence embedding model's performance is indicated by the correlation between the cosine similarity of two sentence vectors and the gold label. The extended STS 2017 dataset consists of three monolingual datasets (en-en, es-es, and ar-ar) and seven multilingual datasets (en-ar, en-de, en-tr, en-es, en-fr, en-it, en-nl).

3.2 Multilingual Sentence Retrieval Task

The multilingual sentence retrieval task discovers the nearest sentence vector for a given query sentence in different languages. The nearest sentence is found by the nearest neighbor using the cosine similarity. The task investigates whether the closest sentence of a query sentence is its translated counterpart.

Tatoeba dataset (Artetxe and Schwenk, 2019) evaluates the multilingual retrieval task for 112 lan-

guages. The dataset contains up to 1,000 English-aligned sentence pairs for each language.

4 Parallel Data

Similar to previous work m-USE (Yang et al., 2019), we consider the following 14 languages as the multilingual setting: ar, de, es, fr, it, ja, ko, nl, pl, pt, ru, th, tr, zh[1].

- **Paracrawl v6.0**: Parallel corpus between English and each of European language from the web (Esplà et al., 2019). We use translation pairs of 8 languages (nl, fr, de, it, pl, pt, es, ru).
- **JParacrawl**: Japanese-English parallel corpus crawled from the web (Morishita et al., 2020).
- **OpenSubtitles 2018**: Parallel corpus among various languages from movie subtitles (Lison and Tiedemann, 2016). We use translation pair between English and each of 14 languages.
- **UN parallel corpus**: Parallel corpus among six languages (ar, en, es, fr, ru, zh) from official records and parliamentary documents of the United Nations (Ziemski et al., 2016). We use parallel corpus between English and each language.
- **SCB En-Thai data**[2]: Thai-English parallel corpus from task-based conversations, organization websites, Wikipedia articles, and government documents.
- **Turkish-parallel-corpora**[3]: Turkish-English parallel corpus from bible, computer application, and website.
- **Korean AIHub data**[4]: Korean-English parallel corpus of literary style and colloquial style. Literary style texts are collected from news, government websites, regulations, and cultural contents. Colloquial style texts are collected scenario-based conversation set.

[1] ar: Arabic, de: German, es: Spanish, fr: French, it: Italian, ja: Japanese, ko: Korean, nl: Dutch, pl: Polish, pt: Portuguese, ru: Russian, th: Thai, tr: Turkish, and zh: Chinese.

[2] https://github.com/vistec-AI/dataset-releases/releases/tag/scb-mt-en-th-2020_v1.0

[3] https://github.com/maidis/turkish-parallel-corpora

[4] http://www.aihub.or.kr/aidata/87/download

4.1 Preprocess

We preprocess the above corpora as follows. All sentences are normalized by NFKC Unicode normalization and collapse diverse double quotations marks, single quotation marks, hyphens, and dashes to a single symbol, respectively. The normalized sentences are tokenized using the SentencePiece tokenizer (Kudo and Richardson, 2018), same as XLM-R. Then translation pairs containing at least one sentence having more than 128 tokens are discarded.

4.2 Data Size

We do not use all of the filtered data but randomly sample to balance the parallel corpus's size among the languages. We randomly sample 2M translation pairs per language. The final multilingual sentence embedding model has seen 28M pairs across all languages in total.

5 Method

The proposed model aims to closely place two sentence vectors from different languages as similar as they are. There are two criteria to accomplish the goal; The first one is to closely place similar sentences in the same language, and the second one is to place sentences with the same meaning in different languages in close proximity to each other. To achieve the goal, we train the model to closely position similar monolingual sentences and then align sentences with the same meaning in different languages.

The training steps start from the XLM-R model (Conneau et al., 2020), which is pre-trained with large monolingual corpora in 100 languages by masked language modeling. For the monolingual training, the XLM-R model is trained on labeled English datasets to learn the similarity between two English sentences. For the multilingual training, the XLM-R model trained in English is extended to the multilingual sentence embedding model by aligning using translation parallel pairs. As the XLM-R model produces embedding vectors for each token, an embedding vector of a given sentence is defined by the mean vector of all embedding tokens.

5.1 Monolingual Training

The goal of monolingual training is to learn the semantic similarity between two sentences of the pivot language. Following the previous work

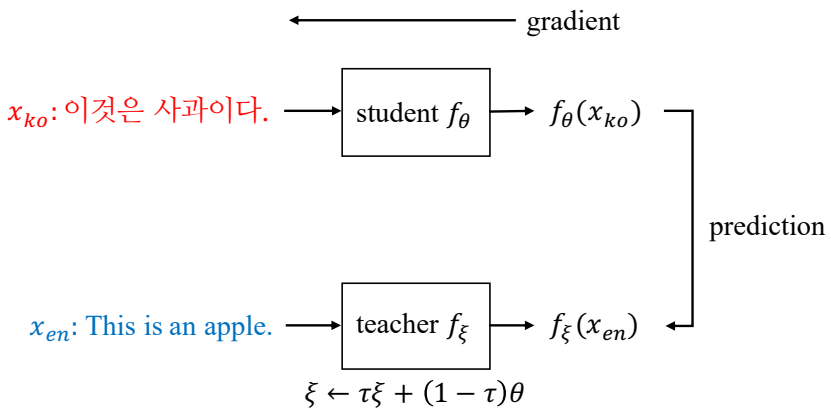

Figure 2: Illustration of multilingual training. In this figure, the pivot language is English, and an English-Korean translation pair (x_{en}, x_{ko}) is given. The student network, which receives the Korean sentence x_{ko} is trained to predict the teacher network's output of English sentence x_{en}. The teacher network is updated by the exponential moving average.

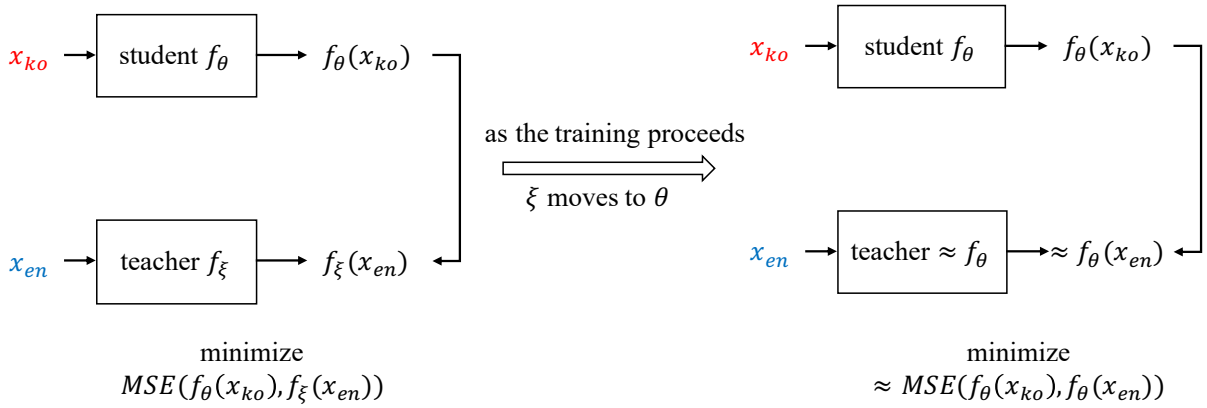

Figure 3: At the beginning of the training process, the student network is trained to minimize the differences between the outputs of the student and the teacher network. As the training proceeds, the parameters of the teacher network are adapted to the student network parameters. Thus, our model can minimize the differences (mean-squared-error) between two sentence inputs in different languages of the student network.

SBERT (Reimers and Gurevych, 2019), we choose XLM-R as the architecture for the monolingual training and fine-tune the architecture using labeled English datasets with siamese networks. The model is firstly trained on SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) with the classification objectives. Then, the model is trained on the STS benchmark (Cer et al., 2017), and its augment (Thakur et al., 2020). The objective is to minimize the mean-squared-error between the cosine similarity and the gold label. In the remaining of the paper, we denote the fine-tuned monolingual model as 'XLM-R-nli-stsb'.

Augmented STS-b dataset is constructed following the method of Augmented SBERT (Thakur et al., 2020). Firstly, while cross-encoder models achieve higher performance than bi-encoder models on the STS benchmark dataset, we train a cross-encoder model from the RoBERTa model (Liu et al., 2019) on the STS benchmark dataset. Secondly, embedding vectors of all sentences from the STS benchmark dataset are constructed using state-of-the-art SBERT model 'stsb-roberta-large'[5]. Thirdly, we find the top 10 nearest sentences of each sentence by nearest neighbor search using Faiss (Johnson et al., 2017). Then, sentence pairs that appeared in the original STS benchmark dataset are discarded. Finally, the cross-encoder labels the similarity of the remaining sentence pairs.

[5] https://github.com/UKPLab/sentence-transformers

5.2 Multilingual Training

After training the sentence embedding model on monolingual similarity dataset in English, the model is aligned across the multilingual sentences using the translation pairs. The alignment of multilingual sentences in a single vector space is done by constructing teacher-student networks. The parameters of the teacher network f_ξ and the student network f_θ are both initialized with the parameters of XLM-R-nli-stsb. However, the two networks are trained by different methods. The overall training procedure is summarized in Figure 2.

The student network is trained to minimize the mean-squared-error between the output sentence vectors of the teacher network and the student network while they receive the pivot language sentence and the corresponding translation pair sentence, respectively. For a given translation pair of the pivot language sentence x_{en} and the corresponding sentence in another language x_{lg}, the loss function for the student network is

$$L = \mathrm{MSE}(f_\theta(x_{lg}), f_\xi(x_{en})). \quad (1)$$

The teacher network, which produces the sentence embedding vectors in the pivot language, is updated by an exponential moving average rather than fixed. For a given decay rate $\tau \in [0, 1]$, the teacher network is updated by

$$\xi \leftarrow \tau\xi + (1 - \tau)\theta \quad (2)$$

for each training step.

Because the teacher network is updated slower than the student network, it produces pivot language sentence embeddings that include more preserved semantics of XLM-R-nli-sts than the student network. Therefore, the teacher network provides the sentence embeddings, which convey the ability to measure the semantic similiarity to the student network.

As the final model used for inference is the student model, the goal of the multilingual training is to align the multilingual sentences of the student network. The loss of the student model Equation 1 seems to align the multilingual sentences between the student and the teacher model. However, as the training proceeds, the parameters of the teacher network gradually move to the parameters of the student network. Thus, at the end of the training steps, the alignment between the multilingual sentences from the student network and the teacher network can be approximated to the alignment between the multilingual sentences of the student network. The explanation is illustrated in Figure 3.

5.3 Training Detail

When we train on multilingual parallel data, we use LARS optimizer (You et al., 2017) with base learning rate 0.02, momentum 0.9, weight decay 1e-6, learning rate warmup over the first 54K steps, and a cosine decay (Loshchilov and Hutter, 2017) of the learning rate with batch size 256. For the teacher network, the exponential moving average parameter τ starts from $\tau_{base} = 0.99999$. Following the settings of BYOL, we set $\tau = 1 - (1 - \tau_{base}) \cdot (\cos(\pi k/K) + 1)/2$ for the current training step k and the maximum number of training steps K while the maximum number of training steps K is set to 105K steps. The training was performed on 8 V100 GPUs and took 3 days. We randomly split 14K parallel sentence pairs for the validation set, and we choose hyperparameters and the final model using it.

6 Experimental Results

In this section, we evaluate the multilingual sentence embedding model on the multilingual similarity evaluation task, the monolingual similarity evaluation task, and the multilingual sentence retrieval task. The first task is the goal of the suggested model, and the latter two tasks are precedent tasks to be successful on the multilingual similarity evaluation task. All tasks include only the test set, not the train/validation set. Also, the proposed model is not trained with any task-specific fine-tuning process. We compare the proposed model to previous multilingual sentence embedding models, m-USE, LASER, LaBSE, and Reimers and Gurevych (2020). Moreover, we did not train the models from the previous papers. The reported scores are from the papers or evaluation using the publicly available models.

6.1 Multilingual Similarity Evaluation Task

As the proposed model aims to compare the similarity between sentences in different languages, we evaluate the model on the multilingual setting of the extended STS 2017 dataset (Reimers and Gurevych, 2020) in Table 1. The suggested model achieves state-of-the-art performance of 84.5. It is 7.5 higher than the XLM-R-nli-stsb model, which is not fine-tuned using multilingual parallel data.

Model	en-ar	en-de	en-tr	en-es	en-fr	en-it	en-nl	Avg.
LASER	66.5	64.2	72.0	57.9	69.1	70.8	68.5	67.0
m-USE	79.3	82.1	75.5	79.6	82.6	84.5	84.1	81.1
LaBSE	74.5	73.8	72.0	65.5	77.0	76.9	75.1	73.5
Reimers and Gurevych (2020)	**82.3**	84.0	80.9	83.1	84.9	86.3	84.5	83.7
XLM-R-nli-stsb mean	62.5	84.2	68.6	77.7	78.3	82.4	85.0	77.0
Our model	76.4	**87.1**	**82.4**	**86.0**	**85.1**	**87.6**	**86.5**	**84.5**

Table 1: Performance on extended STS 2017 similarity evaluation task in the multilingual setting. Scores are reported by 100 × Spearman rank correlation between the cosine similarity of sentence embedding and the gold labels.

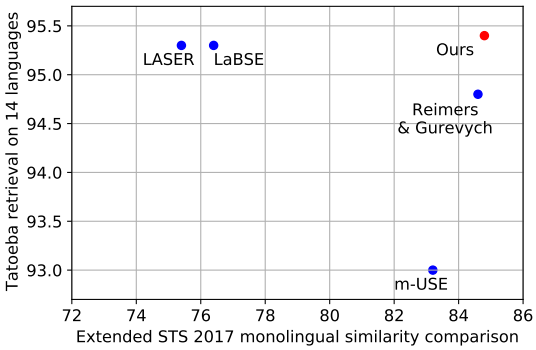

Figure 4: Performances of multilingual sentence embedding models. Previous models, denoted by blue points, solved either the monolingual setting of extended STS 2017 similarity evaluation task or Tatoeba multilingual sentence retrieval task on 14 languages. The suggested model, colored by red, shows the state-of-the-art performance on both tasks.

While the similarity evaluation performance in the pivot language has been preserved in Section 6.2, the improvement comes from the alignment of sentence vectors across the languages as shown in Section 6.3.

The suggested model obtains high performance on the multilingual similarity evaluation task because it is successful in both monolingual similarity evaluation and multilingual sentence retrieval tasks. However, the other models are not successful on the multilingual similarity evaluation task because they succeed on either of one task as shown in Figure 4.

The proposed model performs better than the other models by a significant difference except for Arabic. The proposed model shows consistently poor performance for Arabic in all tasks. We think this is because we executed the same preprocess strategy without any language-specific process for all languages.

6.2 Monolingual Similarity Evaluation Task

The monolingual similarity evaluation task is one of the precedent tasks to be successful on the multilingual similarity evaluation task. The proposed model accomplishes the best score of 84.8 on average on the monolingual setting of the extended STS 2017 dataset as shown in Table 2. Compared to XLM-R-nli-stsb, the scores have been preserved by a slight performance drop.

LASER and LaBSE are unsuccessful on the monolingual similarity evaluation task. LASER and LaBSE focus on separating sentences that do not have the same meaning on the vector space even though they have similar meanings. Because they do not the consider similarity between sentences while they place the sentence vectors, they meet difficulty in capturing semantic similarities.

6.3 Multilingual Sentence Retrieval Task

The multilingual sentence retrieval task is another precedent task for the multilingual similarity evaluation task. We appraise the proposed model on Tatoeba dataset (Artetxe and Schwenk, 2019) to examine the alignment across different languages. Following the language groups of Feng et al. (2020), we group the total 112 languages into four groups. The first 14 languages group is chosen from the language trained by m-USE, which is also selected to train the proposed model. The second 36 languages group is selected by the XTREME benchmark (Hu et al., 2020). The third 82 languages group is the languages that LASER trained. The last group contains the whole sort of languages.

The introduced model achieves the state-of-the-art performance on Tatoeba for 14 languages that we trained. All models we compared are trained on the languages, including the 14 languages.

The model accomplishes second-best performance on the other language groups, even though

Model	en-en	es-es	ar-ar	Avg.
LASER	77.6	79.7	68.9	75.4
m-USE	86.4	86.9	76.4	83.2
LaBSE	79.4	80.8	69.1	76.4
Reimers and Gurevych (2020)	88.8	86.3	**79.6**	84.6
XLM-R-nli-stsb mean	89.8	88.7	77.0	85.2
Our model	**89.3**	**86.4**	78.8	**84.8**

Table 2: Performance on extended STS 2017 similarity evaluation task in the monolingual setting. Scores are reported by 100 $\times$ Spearman rank correlation between the cosine similarity of sentence embedding and the gold labels. We do not sign XLM-R-nli-stsb model's scores to bold because it is not the final proposed model.

Model	14 langs	36 langs	82 langs	All langs
LASER	95.3	84.4	75.9	65.5
m-USE	93.0	44.3	38.5	36.6
LaBSE	95.3	**95.0**	**87.3**	**83.7**
Reimers and Gurevych (2020)	94.8	86.2	75.6	67.0
Our model	**95.4**	89.1	79.4	72.9

Table 3: Performance on Tatoeba sentence retrieval task. Scores are reported by 100 $\times$ accuracy. '14 langs' are languages trained by m-USE and the proposed model. '36 langs' are languages selected by XTREME. '82 langs' are languages trained by LASER.

they contain the languages that are not covered by the model. The proposed model works better than LASER and Reimers and Gurevych (2020), which learned more sort of languages for all the language groups. However, the state-of-the-art model in the language groups that including we do not trained is LaBSE in that they train in 109 languages.

LaBSE tends to distinguish whether each word is present or not, and the suggested model tends to capture the overall meaning. For the query Korean sentence "어리광 부리지 마." (Stop acting like a spoilt child.), while LaBSE chose "Don't be too strict. They're just kids.", the proposed model chose "Don't be ridiculous!" for the most similar sentence. LaBSE seems to choose a sentence containing "kids" because the query sentence contains "어리광" (behave like a spoilt child) which is associate with "child." However, our model seems to catch the whole meaning rather than being associate with a single word.

We also observed the error case in the Afrikaans, which we did not train for the proposed model. For the query sentence "Ek is nou-nou terug met verversings." (I'll be right back with refreshments.), our model choose "I'll get back to you in a moment." because the overall meaning is similar although "refreshments" is missing.

The generalization of multilingual alignment to the untrained language comes from the successful alignment of the proposed model across the languages that have been trained. Experimentally, the proposed model that even trained for only one type of language has also been aligned for the other languages. Before the multilingual training, Tatoeba-14 performance of the model is 89.1, while its Russian score is 91.1 and its Korean score is 85.4. After the multilingual training using only Eng-Rus data for 2M pairs, Tatoeba-14 performance moves to 91.6, while its Russian score is 94.2 and its Korean score is 90.5.

Even though Reimers and Gurevych (2020) is also based on the teacher-student network architecture, they represent insufficient performance on the multilingual sentence retrieval task. The performance drop is caused by their less direct alignment between multiple languages. For example, let the student network f_θ, the teacher network f_ξ, the pivot language sentence x_{en}, and the corresponding sentence in another language x_{ko}. They take intermediate representation $f_\xi(x_{en})$ to align $f_\theta(x_{en})$ and $f_\theta(x_{ko})$ rather than align two student representations directly.

6.4 Decay Rate

The performance of the model can be influenced by the decay rate τ from Equation 2. to check the effect of the decay rate, various settings for the decay rates are examined in Table 4 The optimal

Model	multilingual similarity	monolingual similarity	multilingual retrieval
$\tau = 0.9999$	67.1	68.6	91.1
$\tau = 0.99999$	**84.5**	**84.8**	**95.4**
$\tau = 0.999999$	74.1	78.0	92.4
freeze teacher	74.2	77.7	91.9

Table 4: Performance of various settings of decay rates. Multilingual similarity and monolingual similarity scores are the average on 100 $\times$ Spearman rank correlation of each task. Multilingual retrieval score is the average on 100 $\times$ accuracy of 14 languages trained by the proposed model.

value is empirically selected to 0.99999 and is used for all experiments in the paper.

A small decay rate makes the teacher network forget the previous sentence representation. On the other hand, the teacher network with a large decay rate is not sufficiently similar to the student network. Therefore, the large decay rate hinders the multilingual alignment of the student network which should be learned by aligning between the teacher network and the student network. The frozen teacher network shows a similar result as the large decay rate. Moreover, if the teacher network is updated with the same parameters as the student network, their sentence representations are collapsed to a single vector.

7 Conclusion

This paper introduces the multilingual sentence embedding model, which can compare the similarity between sentences in different languages. The proposed model shows the state-of-the-art performance on the multilingual similarity evaluation task as well as the monolingual similarity evaluation task and the multilingual sentence retrieval task for the languages it has learned. Starting from a model learned to catch similarities between pivot language sentences, the proposed model is trained to align sentence embedding vectors between different languages. The teacher network which produces the pivot language sentence vectors is updated by a slow-moving average rather than fixed. Because the teacher network moves slower than the student network, it conveys sentence embedding vectors which are preserved from the monolingual similarity training. While the teacher network gradually moves to the student network, the alignment between the pivot language sentence of the teacher network and other language sentences of the student network turns into the alignment between the sentence vectors of the student network.

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 R. 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 (EMNLP)*. Association for Computational Linguistics.

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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

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.

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.

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119, Dublin, Ireland. European Association for Machine Translation.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with restarts. *International Conference on Learning Representations*.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

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

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.

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*. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2020. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *arXiv preprint arXiv:2010.08240*.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning?

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

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

Yang You, Igor Gitman, and Boris Ginsburg. 2017. Scaling sgd batch size to 32k for imagenet training. Technical Report UCB/EECS-2017-156, EECS Department, University of California, Berkeley.

Dan Zhang and Anna Khoreva. 2019. Progressive augmentation of gans. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).