Language-agnostic BERT Sentence Embedding

Fangxiaoyu Feng[*], **Yinfei Yang**[*], **Daniel Cer**, **Naveen Arivazhagan**, **Wei Wang**
Google AI
Mountain View
{fangxiaoyu, yinfeiy, cer, navari, wangwe}@google.com

Abstract

We adapt multilingual BERT (Devlin et al., 2019) to produce language-agnostic sentence embeddings for 109 languages. While English sentence embeddings have been obtained by fine-tuning a pretrained BERT model (Reimers and Gurevych, 2019), such models have not been applied to multilingual sentence embeddings. Our model combines masked language model (MLM) and translation language model (TLM) (Conneau and Lample, 2019) pretraining with a translation ranking task using bi-directional dual encoders (Yang et al., 2019a). The resulting multilingual sentence embeddings improve average bi-text retrieval accuracy over 112 languages to 83.7%, well above the 65.5% achieved by the prior state-of-the-art on Tatoeba (Artetxe and Schwenk, 2019b). Our sentence embeddings also establish new state-of-the-art results on BUCC and UN bi-text retrieval.

1 Introduction

Mask language modeling (MLM) pretraining followed by task specific fine-tuning has proven to be a powerful tool for numerous NLP tasks (Devlin et al., 2019). However, pretrained MLMs do not intrinsically produce good sentence-level embeddings. Rather, the production of sentence embeddings from MLMs must be learned via fine-tuning, similar to other downstream task. SentenceBERT (Reimers and Gurevych, 2019) fine-tunes a monolingual BERT based dual-encoder on natural language inference (NLI). The resulting sentence embeddings achieve excellent performance on measures of sentence embedding quality such as the semantic textual similarity (STS) benchmark (Cer et al., 2017) and sentence embedding based transfer learning (Conneau and Kiela, 2018).

[*] Equal contribution.

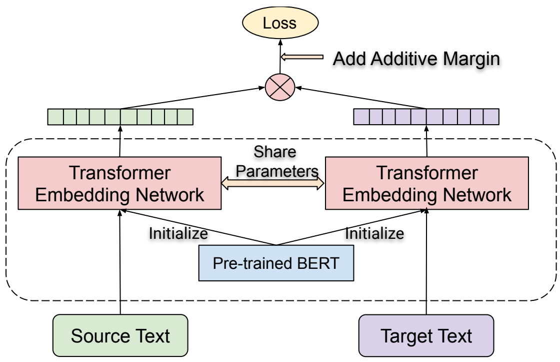

Figure 1: Dual encoder model with BERT based encoding modules.

While multilingual sentence embedding models incorporate similar dual encoders, none explore MLM pretraining. Rather, encoders are trained directly on translation pairs (Artetxe and Schwenk, 2019b; Guo et al., 2018; Yang et al., 2019a), or on translation pairs combined with monolingual input-response prediction (Chidambaram et al., 2019; Yang et al., 2019b). Multilingual sentence embeddings trained directly on translation pairs require large amounts of parallel training data. Moreover, models such as the multilingual universal sentence encoder (*m*-USE) that are trained on multiple languages often perform worse than similar models only targeting a single language pair (Yang et al., 2019a). Paradoxically, multilingual BERT has demonstrated surprisingly good cross-lingual performance without training on parallel translation data (K et al., 2020).

Inspired by these factors, we present a novel method for training multilingual sentence-level embeddings combining existing state-of-the-art methods for multilingual sentence embeddings with MLM and translation language model (TLM) (Conneau and Lample, 2019) pretrained encoders. We employ a dual-encoder framework which consist of paired encoders feeding a combination function. Such models are well suited for

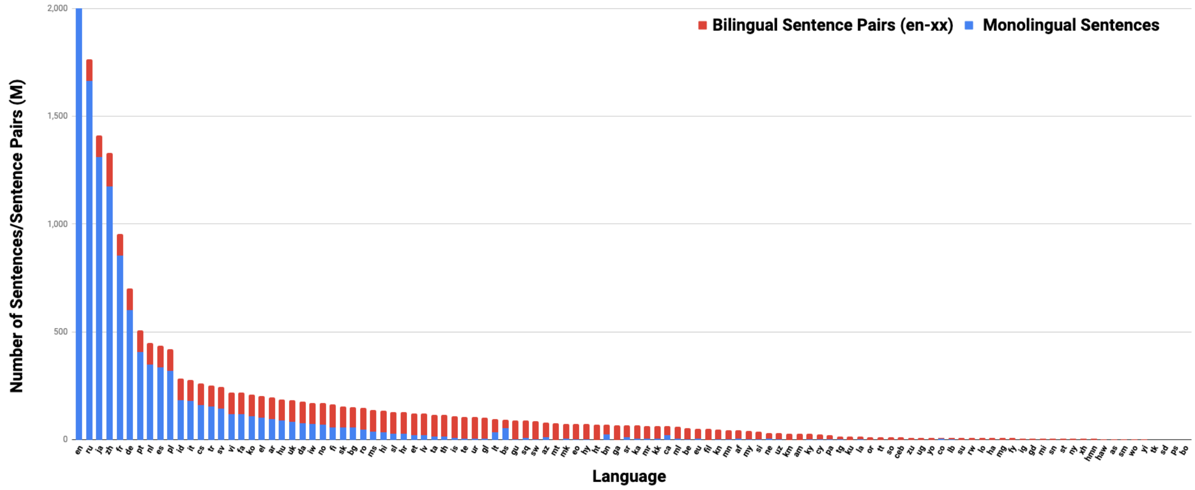

Figure 2: The total size of processed monolingual and bilingual data for all supported languages. Note the en data is way more than the other languages, we cap the maximum value of 2B here.

learning effective cross-lingual sentence embeddings using an additive margin softmax loss (Yang et al., 2019a). Source and target sentences are encoded separately using a shared BERT based encoder (Devlin et al., 2019). The final layer `[CLS]` representations are taken as the sentence embeddings for each input. The similarity between the source and target sentences is scored using cosine over the sentence embeddings produced by the BERT encoders. The model architecture is illustrated in figure 1.

Our multilingual model outperforms the previous state-of-the-art, which are mostly bilingual models, on large bitext mining tasks such as the United Nations corpus (Ziemski et al., 2016) and BUCC (Zweigenbaum et al., 2018), containing pools with tens of millions and hundreds of thousands of translation candidates, respectively. Both tasks cover fr, de, es, ru, and zh, languages all having plenty of training data available. We evaluate the model on the Tatoeba retrieval task (Artetxe and Schwenk, 2019b) covering 112 languages, but with smaller pools of only between 100 to 1000 translation candidates. Compare to LASER (Artetxe and Schwenk, 2019b), the new model achieves matching performance on languages with plenty of training data, but does significantly better on languages with limited data, boosting the averaged accuracy on the entire 112 language evaluation to 83.7% from the 65.5% achieved by the previous state-of-art. We also observe the model performs strongly on the Tatoeba task for 30+ languages for which we have no monolingual or bilingual training data. We believe this is achieved by large scale training on 109 languages, covering language families containing the unseen languages.

The novel contributions of this paper are: (1) A combination of pre-training and finetuning strategies to boost the performance of a dual encoder translation ranking model to the state-of-the-art performance on bi-text mining; (2) A single massively multilingual model spanning 109 languages and showing cross-lingual transfer even to zero-shot cases. (3) A thorough analysis and ablation study to understand the impact of various data quality, data quantity, pre-training and negative sampling strategies. Our model is available at `https://tfhub.dev/google/LaBSE`

2 Corpus

We have two types of data: monolingual data and bilingual translation pairs.

Monolingual Data We collect monolingual data from CommonCrawl[1] and Wikipedia[2]. We use the 2019-35 version of CommonCrawl with heuristics from Raffel et al. (2019) to remove noisy text. Additionally, we remove short lines < 10 characters and those > 5000 characters.[3] The wiki data is extracted from the 05-21-2020 dump using WikiExtractor[4]. An in-house tool splits the text into sentences. The sentences are filtered using a sentence

[1] `https://commoncrawl.org/`

[2] `https://www.wikipedia.org/`

[3] Long lines are usually JavaScript or attempts at SEO.

[4] `https://github.com/attardi/wikiextractor`

做了一个
句子质量
分类器

quality classifier. The quality classifier is trained using sentences from the main content of webpages as positives and text from other areas as negatives. After filtering, we obtain 17B monolingual sentences, about 50% of the unfiltered version.

Bilingual Translation Pairs The translation corpus is constructed from the web pages using a bitext mining system similar to the approach described in Uszkoreit et al. (2010). The extracted sentence pairs are filtered by a pre-trained contrastive-data-selection (CDS) scoring model (Wang et al., 2018). Human annotators manually evaluate sentence pairs from a small subset of the harvested pairs and mark the pairs as either GOOD or BAD translations. The data-selection scoring model threshold is chosen such that 80% of the retrained pairs from the manual evaluation are rated as GOOD. We further limit the maximum sentence pairs to 100 million for each language to balance the data distribution. Many languages still have far fewer than 100M sentences. The final corpus contains 6B translation pairs. [5] The distribution for each language is shown in figure 2.

3 Models

Dual-encoders contain paired encoders feeding a scoring function, as in figure 1. The source and target sentences are encoded separately. Sentence embeddings are extracted from the last hidden state of the encoder `[CLS]` token.

3.1 Bidirectional Dual Encoder with Additive Margin Softmax

Following Yang et al. (2019a), we train bidirectional dual encoders with additive margin softmax loss with in-batch negative sampling:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \frac{e^{\phi(x_i,y_i)-m}}{e^{\phi(x_i,y_i)-m} + \sum_{n=1,n\neq i}^{N} e^{\phi(x_i,y_n)}} \tag{1}$$

The embedding space similarity of x and y is given by $\phi(x,y)$. Following prior work (Yang et al., 2019a), we set $\phi(x,y) = cosine(x,y)$. The loss attempts to rank y_i, the true translation of x_i, over all $N-1$ alternatives in the same batch even when $\phi(x_i,y_i)$ is discounted by margin m. Notice that $\mathcal{L}$ is asymmetric and depends on whether the softmax is over the source or the target. To bi-directional ranking, the final loss function sums the source to target, $\mathcal{L}$, and target to source, $\mathcal{L}'$, losses:

$$\bar{\mathcal{L}} = \mathcal{L} + \mathcal{L}' \tag{2}$$

3.2 Cross-Accelerator Negative Sampling

Cross-lingual embedding models trained with in-batch negative samples benefits from large training batch sizes (Guo et al., 2018). Resource intensive models like BERT, are limited to small batch sizes due to memory contraints. While data-parallelism allows us to increase the effective batch size by using multiple accelerators, the batch-size on an individual core does not change the batch over multiple accelerators. However, this results in a smaller local batch size on each accelerator. For example, with a per-core batch size of 128, each example only receives 127 negative examples.

一个batch size
中一个是正样本,
其他都是负样本

We introduce *cross-accelerator negative sampling*.[6] As illustrated in figure 3, under this strategy the sentences from all cores are broadcast as negatives for the examples assigned to other cores. This allows us to fully realize the benefits of distributed training.

3.3 Pre-training and parameter sharing

We use a transformer encoder (Vaswani et al., 2017). The encoder is pre-trained with Masked Language Model (MLM) (Devlin et al., 2019) and Translation Language Model (TLM)Conneau and Lample (2019)[7] (Conneau and Lample, 2019) training on the monolingual data and bilingual translation pairs, respectively. For an L layer transformer encoder, we train using a 3 stage progressive stacking algorithm (Gong et al., 2019), where we first learn a $\frac{L}{4}$ layers model and then $\frac{L}{2}$ layers and finally all L layers. The parameters of the models learned in the earlier stages are copied to the models for the subsequent stages.

4 Evaluation

4.1 Configuration

In this section we describe the training details for the dual encoder model. We employ the wordpiece

[5] While we have a large bilingual dataset to work with, experiments in later sections that show even 200M pairs across all languages is sufficient.

[6] While our experiments use TPU accelerators, the same strategy can also be applied to models trained on GPU.

[7] Diverging from Conneau and Lample (2019), we do not provide a language hint to encourage multilinguality.

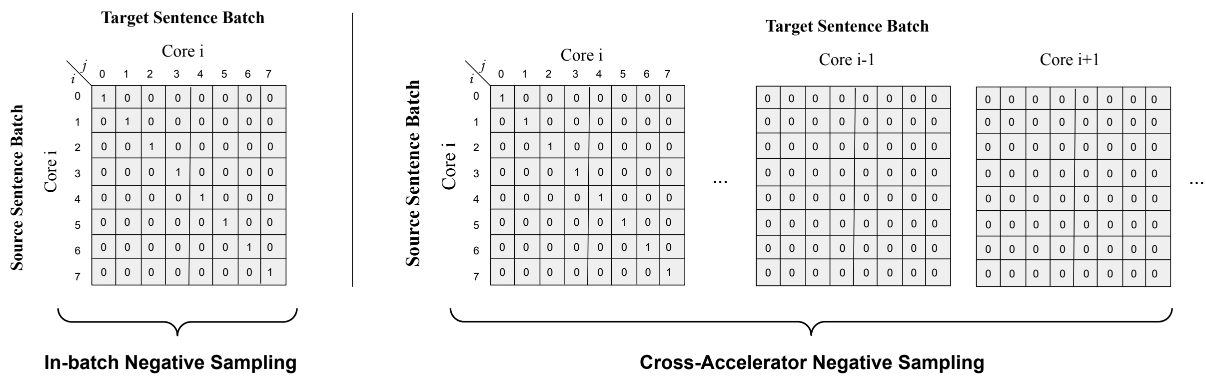

Figure 3: Negative sampling example in a dual encoder framework. The dot-product scoring function makes it efficient to compute the pairwise scores in the same batch with matrix multiplication. The value in the grids indicates the ground truth labels, with all positive labels located in diagonal grids. **[Left]**: The in-batch negative sampling in a single core; **[Right]**: *Synchronized multi-accelerator negative sampling* using n TPU cores and batch size 8 per core with examples from other cores are all treated as negatives.

model (Sennrich et al., 2016) to tokenize the text input. A new cased vocabulary is built from the all data sources using the wordpiece vocabulary generation library from Tensorflow Text[8]. The language smoothing exponent from the vocab generation tool is set to 0.3, as the distribution of data size for each language is imbalanced. The final vocabulary size is 501,153.

The encoder architecture follows the BERT Base model which uses 12 layers transformer with 12 heads and 768 hidden size. The encoder parameters are shared for all languages. We take the [CLS] token representation from the last layer as the sentence embeddings, The final embeddings are l_2 normalized.

The pre-trained BERT model is trained on 512-core slices of Cloud TPU V3 pods using a batch size 8192. The max sequence length is set to 512 and 20% of tokens (or 80 tokens at most) per sequence are masked the MLM and TLM predictions. We train 400k, 800k, 1.8M steps for each stage using all the monolingual and bilingual data.

The LaBSE models are trained on 32-core slices of Cloud TPU V3 pods using a global batch size 2048 with max sequence length 64 for both of the source and target. Following Yang et al. (2019a), a margin value of 0.3 is used in all experiments. The final models are trained 50K steps (less than 1 epoch) using AdamW optimizer with initial learning rate 1e-5 and linear weight decay. Note that the models only see 200 million parallel pairs by training 50K steps. During training, the sentence embeddings (after normalization) are multiplied by a scaling factor following Chidambaram et al. (2018), we set the scaling factor to 10. We observe that the scaling factor is important for training a dual encoder model with the normalized embeddings. All parameters are tuned on a hold-out development set.

4.2 BUCC

The BUCC mining task is a shared task on parallel sentence extraction from two monolingual corpora with a subset of them assumed to be parallel, and that has been available since 2016. We make use of the data from the 2018 shared task, which consists of corpora for four language pairs: fr-en, de-en, ru-en and zh-en. For each language pair, the shared task provides a monolingual corpus for each language and a gold mapping list containing true translation pairs. These pairs are the ground truth. The task is to construct a list of translation pairs from the monolingual corpora. The constructed list is compared to the ground truth, and evaluated in terms of the F1 measure. For more details on this task refer to (Zweigenbaum et al., 2018).

Original BUCC task has train test split and the test ground truth are blinded. Recent cross-lingual retrieval work evaluate on the train set with best F1 without using any of the in-domain data (Yang et al., 2019b; Hu et al., 2020). We follow the setup to report the best retrieval performance on train set using the raw cosine similarity score of the LaBSE

[8] https://github.com/tensorflow/text/blob/master/tools/wordpiece_vocab/generate_vocab.py

	Models	fr-en			de-en			ru-en			zh-en		
		P	R	F	P	R	F	P	R	F	P	R	F
Forward	Artetxe and Schwenk (2019a)	82.1	74.2	78.0	78.9	75.1	77.0	-	-	-	-	-	-
	Yang et al. (2019a)	**86.7**	85.6	86.1	90.3	88.0	89.2	84.6	91.1	87.7	86.7	**90.9**	88.8
	LaBSE	86.6	**90.9**	**88.7**	**92.3**	**92.7**	**92.5**	**86.1**	**91.9**	**88.9**	**88.2**	89.7	**88.9**
Backward	Artetxe and Schwenk (2019a)	77.2	72.7	74.7	79.0	73.1	75.9	-	-	-	-	-	-
	Yang et al. (2019a)	83.8	85.5	84.6	89.3	87.7	88.5	83.6	90.5	86.9	**88.7**	87.5	88.1
	LaBSE	**87.1**	**88.4**	**87.8**	**91.3**	**92.7**	**92.0**	**86.3**	**90.7**	**88.4**	87.8	**90.3**	**89.0**

Table 1: [P]recision, [R]ecall and [F]-score of BUCC training set score with cosine similarity scores. The thresholds are chosen for the best F scores on the training set. Following the naming of BUCC task (Zweigenbaum et al., 2018), we treat en as the target and the other language as source in forward search. Backward is vice versa.

embeddings[9].

Table 1 shows the BUCC performance of proposed model comparing with two baselines from Artetxe and Schwenk (2019a) and Yang et al. (2019a). Following the original previous work, we perform both of the forward search and backward search. Where forward search treat en as the target and the other language as source in forward search, and backward is vice versa. The LaBSE outperforms the previous models in all languages. It is worth to note that the previous state-of-the-art (Yang et al., 2019a) are bilingual models, while LaBSE covers 109 languages.

4.3 United Nations

We then evaluate the the United Nations Parallel Corpus (Ziemski et al., 2016), which consists of 86,000 bilingual document pairs in five language pairs: from en to fr, es, ru, ar and zh. Document pairs are near perfectly aligned, total of 11.3 million[10] aligned sentence pairs can be parsed from the document pairs. As noticed in Guo et al. (2018), strong models can easily reach perfect performance if the candidate size is small, this dataset is good to differentiate those models with its large candidate sets.

For each non-English language, we iterate over English sentences to find the translation sentence from the entire sentence pool from the other language. Table 2 shows precision@1 (P@1) for the experimented models. We compare the proposed model with the current state-of-the-art bilingual models from Yang et al. (2019a) and public multilingual universal sentence encoder (*m*-USE) model with the transformer architecture.

Model	Langs	en-es	en-fr	en-ru	en-zh
Yang et al. (2019a)	2	89.0	86.1	89.2	**87.9**
m-USE$_{\text{Trans.}}$	16	86.1	83.3	88.9	78.8
LaBSE	109	**91.1**	**88.3**	**90.8**	87.7

Table 2: P@1 on UN parallel sentence retrieval task.

Again LaBSE shows the new state-of-the-art performance on 3 of the 4 languages, with P@1 91.1, 88.3, 90.8 for en-es, en-fr, en-ru respectively. The zh-en of LaBSE reaches 87.7, only 0.2 lower than the best bilingual model and nearly 9 points better than the previous best multilingual model.

4.4 Tatoeba

To better assess the performance on massive languages, we further evaluate the proposed model on the Tatoeba corpus introduced by Artetxe and Schwenk (2019b). This dataset consists up to 1,000 English-aligned sentence pairs for 112 languages and the task is finding the nearest neighbor for each sentence in the other language using cosine similarity distance. The accuracy for each language and average accuracy are computed.

We evaluate performance on several groupings of languages for fair-comparison and to identify broader trends. The first 14 language group is selected from the languages covered by *m*-USE. We also evaluate the second language group with 36 languages from the XTREME benchmark (Hu et al., 2020). The third 82 language group, selected from the languages that LASER has training data, should covers some tail languages. At last, we compute the average accuracy for all languages.

Table 3 shows the macro-average accuracy of different language groups for the LaBSE, comparing against *m*-USE and LASER. As expected, all these models perform strongly on the 14 languages group that covers most head languages,

[9] Note a second stage scoring model can be applied to get improved performance, e.g. margin based scorer (Artetxe and Schwenk, 2019a), BERT fine-tuning classifier (Yang et al., 2019a). We treat it as an independent work.

[10] About 9.5 million after de-duping.

Model	14 Langs	36 Langs	82 Langs	All Langs
m-USE$_{\text{Trans.}}$	93.9	–	–	–
LASER	**95.3**	84.4	75.9	65.5
LaBSE	**95.3**	**95.0**	**87.3**	**83.7**

Table 3: Accuracy(%) of the Tatoeba datasets. **[14 Langs]**: The languages USE supports. **[36 Langs]**: The languages selected by XTREME. **[82 Langs]**: Languages that LASER has training data. **All Langs**: All languages supported by Taoteba.

with >93% average accuracy for all there models. LASER and LaBSE are slightly better than *m*-USE. By including more languages, the averaged accuracy for both of LASER and LaBSE become lower. LaBSE starts outperforming LASER more with including more languages, with +10.6%, +11.4%, and +18.2% average accuracy better on all 36 languages, 82 languages, and 112 languages respectively.

5 Analysis

Additive Margin The additive margin (Yang et al., 2019a) is still an very important part for learning the effective cross-lingual embedding space. The improvement on the large scale UN retrieval task is very large against the base model even with a very small margin value, as shown in row 5-7 of table 4, The model with additive margin value 0 perform poorly on all 4 UN languages with 60s or 70s P@1. With a small margin value 0.1 the model is improved significantly with 80+ P@1 for all languages. The models with margin 0.2 and margin 0.3 (the finale model) have the similar ballpark performance, with margin=0.2 model performs slightly better on Tatoeba and margin=0.3 model performs better on UN and BUCC. We select the margin 0.3 as we observe the evaluation the larger scale tasks are more stable from preliminary experiments.

Pre-training. We first experiment the model without BERT pre-training. The results are listed in row 1-4 of table 4. The model trained with default training steps, e.g. 50K, perform poorly comparing with the model with pre-training. We further train the model with longer steps including 100K, 200K, until 500K. The performance keep increasing and approaching to the model with pre-training around 500k steps. The overall performance is still slightly worse, keep training doesn't increase the model performance significantly. The model will see 1B examples with training 500K steps, while the 50K model only sees the 200M examples [11]. Indicating that the pre-training also leads to a significant less requirement of the parallel training data.

Comparison to Multilingual BERT We compare out pre-training approach against initializing from multilingual BERT model[12]. This model perform strongly on the head languages in UN and BUCC tasks, with high 80s P@1 and best F1 for all UN and BUCC languages respectively. However, it perform significantly underperforms on Tatoeba tasks, with -2.8 average accuracy on the 36 language set and average accuracy on the all language set.

Our pre-training approach improves over multilingual BERT on tail languages due to a combination of reasons. We use a much larger vocab , 500k versus 30K, which has been shown to improve multilngual performance (Conneau et al., 2019). We also include TLM in addition to MLM as this has been shown to improve cross lingual transferConneau and Lample (2019). Finally, we pretrain on common crawl which is much larger, albeit noiser, than the wiki data multinlgual BERT is train on.

Importance of the Data Selection The LaBSE models are trained with the data that selected by a pre-trained contrastive data selection (CDS) model. In order to understand how the data selection affect the model performance, we also train a model with the original web crawled translation pairs without CDS selection, which are still good enough to train reasonably good NMT models. Surprisingly, this model doesn't perform well on the retrieval task even from a 100 candidates pool with around 80% precision@1, comparing against 99% precision@1 of a model trained with clean data. The result indicates that the dual encoder model training is sensitive to the data quality. Note that the CDS selection is not only based on the quality but also based on a domain match with the training data (Wang et al., 2018), so that the selected data could possibly falls into a narrow domain where is the CDS training data from. A dedicated translation quality model could improve the data selection stage further or increase the coverage, we leave this as a future work.

[11] It is relative easy to get 200M parallel examples for many languages from public sources like Paracrawl, TED58.

[12] multi_cased_L-12_H-768_A-12

Model	Control Variables	Training Steps	UN (en → xx)				BUCC (xx → en)				Taoeba (xx → en)	
			es	fr	ru	zh	fr	de	ru	zh	36 Langs	All
LaBSE	no pretrain	50K	83.6	75.6	75.8	70.4	–	–	–	–	–	–
LaBSE	no pretrain	100K	86.5	82.2	82.8	79.9	–	–	–	–	–	–
LaBSE	no pretrain	200K	89.1	85.3	86.8	83.0	–	–	–	–	–	–
LaBSE	no pretrain	500K	90.0	87.3	89.8	85.2	88.3	92.0	88.6	85.8	94.8	82.4
LaBSE	margin = 0	50K	73.7	62.2	64.4	79.2	–	–	–	–	–	–
LaBSE	margin = 0.1	50K	88.0	82.7	86.8	83.9	–	–	–	–	–	–
LaBSE	margin = 0.2	50K	90.2	87.8	89.7	87.2	87.9	91.9	88.6	88.2	**95.2**	**83.9**
LaBSE	init. mBERT	50K	89.3	85.7	89.3	87.2	86.8	90.5	87.3	87.4	92.2	78.4
LaBSE	full model	50K	**91.1**	**88.3**	**90.8**	**87.7**	**88.7**	**92.7**	**88.9**	**88.9**	95.0	83.7

Table 4: UN (P@1), BUCC (F score from forward search), and Taoteba (Average accuracy) performance for different model configurations in ablation study. The full model is initialized from the customized BERT model, using margin value 0.3, and trained for 50K steps.

5.1 Zero-shot Transfer to Languages without Training Data

Figure 4 list the Tatoeba accuracy for those languages we don't have any training data. There are total of 30+ such languages[13]. The performance is surprisingly good for most of the languages with an average accuracy around 60% on those languages. Nearly one third of them have accuracy larger than 75%, and only 7 of them have accuracy lower than 25%. Such positive language transfer across languages is only possible due to the massively multilingual nature of LaBSE. The top and bottom 5 of those languages languages are listed in table 5. We analyze effect of vocabulary by inspecting the unknown token rate, average token length in characters, and average sentence length in tokens of all languages (Arivazhagan et al., 2019). The unknown token rates are surprisingly low for all languages, indicating a good coverage of the built vocab. Languages with low performance tend to on average have shorter token lengths and longer sequence lengths indicating which indicates sub-par vocab coverage . A better vocab could potentially benefit those languages.

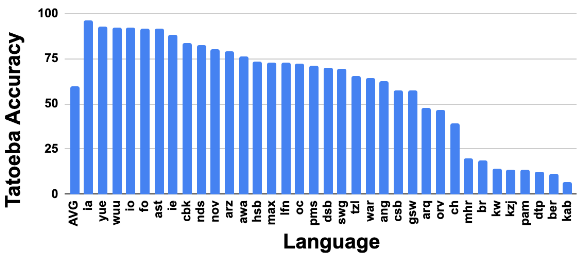

Figure 4: Tatoeba accuracy for those languages without any training data. The average (AVG) accuracy is 64.1%, listed at the first.

Language	Tatoeba Accuracy	Unknown Token %	Avg. Token. Length	Avg. Sent. Length
nb	98.9	0	3.47	8.29
tl	97.3	0	3.81	8.57
ia	96.3	0	3.58	11.65
pes	96	0	3.11	9.20
he	93.6	0	3.04	8.63
kzj	13.5	0	3.30	12.66
pam	13.5	0	3.17	10.82
dtp	12.6	0	3.26	11.14
ber	11.3	0.01	2.82	12.99
kab	6.8	0.01	2.74	12.72

Table 5: The top and bottom performance of Tatoeba for languages without training data.

Negative Sampling Here we measure the impact of cross-accelerator negative sampling. We also briefly explore using hard rather than random negatives. Results are shown in table 6.

Guo et al. (2018) explored hard negative mining in the dual encoder framework for learning cross-lingual embeddings, and this technology has been used as the default setup for followup works (Chidambaram et al., 2018; Yang et al., 2019a). We experiment the hard negative mining for Spanish (es) following Guo et al. (2018) within this LaBSE setup. A weaker dual encoder using the deep averaging network is trained to mine the negatives from the bilingual pool of en-es. Similar to the cross-accelerator negatives, the mined negatives are also appended to each example. Due to the memory constraint, we only append 3 mined hard negatives in es languages for each en source sentence. Since the amount of examples increased 4x

[13] Language mapping is done manually, some languages are close to those languages with training data but may be treated differently according to ISO-639 standards and other information.

Model	es	fr	ru	zh	avg.
base model	91.1	88.3	90.8	87.7	89.5
no cross-accelerator sampling	90.3	87.9	91.1	86.6	89.0
w/ es hard negatives	90.4	87.1	89.9	87.2	88.7

Table 6: P@1 on UN with different negative sampling strategies.

Model	dev	test
SentenceBERT (Reimers and Gurevych, 2019)	-	79.2
m-USE (Yang et al., 2019b)	**83.7**	**82.5**
USE (Cer et al., 2018)	80.2	76.6
ConvEmbed (Yang et al., 2018)	81.4	78.2
InferSent (Conneau et al., 2017)	80.1	75.6
LaBSE	74.3	72.8
STS Benchmark Tuned		
SentenceBERT-STS (Reimers and Gurevych, 2019)	-	86.1
ConvEmbed (Yang et al., 2018)	83.5	80.8

Table 7: Semantic Textual Similarity (STS) benchmark(Cer et al., 2017) performance as measured by Pearson's r.

per en sentence in es batches, We also decrease batch size from 128 to 32 in the hard negative experiment. To get fair comparison, we trained 200k steps for the experiment instead of 50k in other experiments. For languages other than es, the training data was same as other experiments but batch size was decreased to 32 together.

Removing the cross-accelerator sampling hurts the model performance slightly on average. We also didn't see hard negatives help the LaBSE model's performance. The P@1 for es is 90.4 while the full model is 91.1. Due to the impact of batch size, performance of other major languages also decrease.

5.2 Semantic Similarity

The Semantic Textual Similarity (STS) benchmark (Cer et al., 2017) measures the ability of models to replicate fine-grained graded human judgements of pairwise English sentence similarity. Models are scored according to their Pearson correlation, r, on gold labels ranging from 0, unrelated meaning, to 5, semantically equivalent, with intermediate values capturing carefully defined degrees of meaning overlap. STS is broadly used to evaluate the quality of sentence-level embeddings by assessing the degree to which similarity between pairs of sentence embeddings aligns with human perception of sentence meaning similarity.

Table 7 reports performance on the STS benchmark for LaBSE versus existing sentence embedding models. Following prior work, the semantic similarity of a sentence pair according to LaBSE is computed as the arccos distance between the pair's sentence embeddings.[14] In addition to performance using similarity scores commuted directly from sentence embeddings from the various models, we include numbers for SentenceBERT when it is fine-tuned for the STS task as well as ConvEmbed when an additional affine transform is trained to fit the embeddings to STS. Rather than directly measuring sentence embeddings quality, tuning toward STS informs the extent to which the information required to assess semantic similarity is captured in anyway within the model.

We observe that LaBSE performs worse on pairwise English semantic similarity than other sentence embedding models. This result contrasts with its excellent performance on cross-lingual bi-text retrieval. The cross-lingual *m*-USE model notably achieves the best overall performance, even outperforming SentenceBERT when SentenceBERT is not fine-tuned for the STS task. We suspect training LaBSE on translation pairs biases the model to excel at detecting meaning equivalence, but not at distinguishing between fine grained degrees of meaning overlap. *m*-USE is trained similarity to LaBSE but also contains additional monolingual training data on the prediction of input-response pairs. Predicting input-response pairs has been previously shown to produce excellent sentence embedding representations as assessed by semantic similarity tasks (e.g., ConvEmbed (Yang et al., 2018)).

6 Mining Parallel Text from CommonCrawl

We employ the LaBSE model to mine parallel text from CommonCrawl, a large-scale monolingual corpus, and train NMT models on the mined data. We experiment with two language pairs: English-Chinese (en-zh) and English-German (en-de). The processed CommonCrawl corpus explained in section 2 is used. There are total 1.3B, 0.7B, 7.7B sentences after processing for zh, de, and en respectively.

For each of the language pairs, we treat sentences in one language as source, and sentences in the other language as target. The dual-encoder model can easily encode the source and target sen-

[14] Within prior work, *m*-USE, USE and ConvEmbed use arccos distance to measure embedding space semantic similarity, while InferSent and SentenceBERT use cosine similarity.

tences separately. Taking advantage of this property, we first pre-encode all target sentences into a target database, and then we iterate through the source sentences to retrieve the potential targets for each one of them using an approximated nearest neighbour (ANN) search (Vanderkam et al., 2013), which is sub-linear with respect to the target database size. Given the fact that en data is almost 10x larger than the other languages, we use the zh and de sentences to retrieve the indexed en sentences, to be more efficient.

After the retrieval step, every source sentence can be paired with its nearest neighbors. To filtering those pairs that are absolute not translation of each other, we keep those pairs with similarity scores ≥ 0.6 only[15]. There are 261M and 104M such sentences for en-zh and en-de, respectively[16].

For en-de or en-zh, we train a model with Transformer-Big (Vaswani et al., 2017) in the following way: First we train the model on the mined data as is for 120k steps with batch size 10k on TPU. Then we select top-20% with a data selection method (Wang et al., 2018), and train for another 80k steps. And then we evaluate the final model. We carry out the second step because, after examination, we notice there are better-quality sentence pairs than the overall mined data on average and thus can be selected to make the system better.

Results in table 8 show the strength of the mined data and indicate the headroom. By comparing with a previous en-de result (Edunov et al., 2018), we see the mined data yields performance that is 3 BLEU away from WMT17 en-de parallel data. By comparing with a previous en-zh result (Sennrich et al., 2017), we see that the model is 0.6 BLEU away from a WMT17 NMT model (Sennrich et al., 2017) that is trained on the WMT parallel data. This indicates our headroom by tuning the method and mining more data.

[15]The threshold 0.6 is selected by manually inspect the data, that the pairs greater or equal to this threshold are likely to be translation or partial translation of each other. Note the pairs could still be noisy, we relay on the data selection step described below to select clean sentence pairs for training NMT models.

[16]Due the time and resource constraints, only roughly 40 percents of the source sentences for each language pair are processed to mine potential translations.

Language	# of XX Sents	# of En Sents	# of Mined Pairs	BLEU (en→xx)
en-zh	560M	7.7B	261M	35.7
en-de	330M	7.7B	104M	27.2

Table 8: The number of source / target sentences and number of mined parallel text from CommonCrawl. BLEU scores are evalauted on wmtnews17 and wmtnews14 for zh-en and de-en respectively.

7 Conclusion

This paper presents a language-agnostic BERT sentence embedding model supporting 109 languages. We introduce a simple approach to adopt a pre-trained BERT model to dual encoder model to train the cross-lingual embedding space effectively and efficiently. The model achieves the state-of-the-art performance on various bitext retrieval/mining tasks compareing with previous state-of-the-art with less language coverage. We also show the model performs strongly even on those languages LaBSE doesn't have any training data, as long as the text can be segment to the wordpiece tokens reasonably. Extensive experiments show the additive margin softmax is a key factor for training the model, parallel data quality matters, but the amount parallel data required could be largely diminished with the masked language model pre-training. The pre-trained model is released at tfhub to support further research on this direction and possible downstream applications.

Acknowledgments

We thank our teammates from Descartes, Translate and other Google groups for their feedback and suggestions. Special thanks goes to Sidharth Mudgal, and Jax Law for help with data processing; as well as Jialu Liu, Tianqi Liu, Chen Chen, and Anosh Raj for help on BERT pretraining.

References

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Mikel Artetxe and Holger Schwenk. 2019a.

Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019b. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Linguistics*, 7:597–610.

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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. 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.

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

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

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 7059–7069. Curran Associates, Inc.

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

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. Association for Computational Linguistics.

Linyuan Gong, Di He, Zhuohan Li, Tao Qin, Liwei Wang, and Tieyan Liu. 2019. Efficient training of BERT by progressively stacking. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages

2337–2346, Long Beach, California, USA. PMLR.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP*.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The university of Edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399, Copenhagen, Denmark. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1101–1109, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dan Vanderkam, Rob Schonberger, Henry Rowley, and Sanjiv Kumar. 2013. Nearest neighbor search in google correlate. Technical report, Google.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. 2018. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation*, pages 133–143. Association for Computational Linguistics.

Yinfei Yang, Gustavo Hernández Ábrego, Steve Yuan, Mandy Guo, Qinlan Shen, Daniel Cer, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019a. Improving multilingual sentence embedding using bi-directional dual encoder with additive margin softmax. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5370–5378. ijcai.org.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019b. Multilingual universal sentence encoder for semantic retrieval. *CoRR*, abs/1907.04307.

Yinfei Yang, Steve Yuan, Daniel Cer, Shengyi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 164–174, Melbourne, Australia. Association for Computational Linguistics.

Michal 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. European Language Resources Association (ELRA).

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2018. Overview of the third bucc shared task: Spotting parallel sentences in comparable corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

ISO	NAME	ISO	NAME	ISO	NAME
af	AFRIKAANS	ht	HAITIAN_CREOLE	pt	PORTUGUESE
am	AMHARIC	hu	HUNGARIAN	ro	ROMANIAN
ar	ARABIC	hy	ARMENIAN	ru	RUSSIAN
as	ASSAMESE	id	INDONESIAN	rw	KINYARWANDA
az	AZERBAIJANI	ig	IGBO	si	SINHALESE
be	BELARUSIAN	is	ICELANDIC	sk	SLOVAK
bg	BULGARIAN	it	ITALIAN	sl	SLOVENIAN
bn	BENGALI	ja	Japanese	sm	SAMOAN
bo	TIBETAN	jv	JAVANESE	sn	SHONA
bs	BOSNIAN	ka	GEORGIAN	so	SOMALI
ca	CATALAN	kk	KAZAKH	sq	ALBANIAN
ceb	CEBUANO	km	KHMER	sr	SERBIAN
co	CORSICAN	kn	KANNADA	st	SESOTHO
cs	CZECH	ko	KOREAN	su	SUNDANESE
cy	WELSH	ku	KURDISH	sv	SWEDISH
da	DANISH	ky	KYRGYZ	sw	SWAHILI
de	GERMAN	la	LATIN	ta	TAMIL
el	GREEK	lb	LUXEMBOURGISH	te	TELUGU
en	ENGLISH	lo	LAOTHIAN	tg	TAJIK
eo	ESPERANTO	lt	LITHUANIAN	th	THAI
es	SPANISH	lv	LATVIAN	tk	TURKMEN
et	ESTONIAN	mg	MALAGASY	tl	TAGALOG
eu	BASQUE	mi	MAORI	tr	TURKISH
fa	PERSIAN	mk	MACEDONIAN	tt	TATAR
fi	FINNISH	ml	MALAYALAM	ug	UIGHUR
fr	FRENCH	mn	MONGOLIAN	uk	UKRAINIAN
fy	FRISIAN	mr	MARATHI	ur	URDU
ga	IRISH	ms	MALAY	uz	UZBEK
gd	SCOTS_GAELIC	mt	MALTESE	vi	VIETNAMESE
gl	GALICIAN	my	BURMESE	wo	WOLOF
gu	GUJARATI	ne	NEPALI	xh	XHOSA
ha	HAUSA	nl	DUTCH	yi	YIDDISH
haw	HAWAIIAN	no	NORWEGIAN	yo	YORUBA
he	HEBREW	ny	NYANJA	zh	Chinese
hi	HINDI	or	ORIYA	zu	ZULU
hmn	HMONG	pa	PUNJABI		
hr	CROATIAN	pl	POLISH		

Table 9: The supported languages of LaBSE (ISO 639-1/639-2).