1.论文本身是为了做一个sentence embedding，只不过在这个过程中增加了多语言的任务，从单语言到多语言的过程运用了蒸馏的手段
2.训练语料都除了原始句子还有其翻译。
3.附带翻译的原因是，论文的前提是基于一个假设，就是原始句子和其翻译在经过model后，会被映射到同一个空间。

Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation

Nils Reimers and Iryna Gurevych
Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
www.ukp.tu-darmstadt.de

Abstract

We present an easy and efficient method to extend existing sentence embedding models to new languages. This allows to create multilingual versions from previously monolingual models. The training is based on the idea that a translated sentence should be mapped to the same location in the vector space as the original sentence. We use the original (monolingual) model to generate sentence embeddings for the source language and then train a new system on translated sentences to mimic the original model. Compared to other methods for training multilingual sentence embeddings, this approach has several advantages: It is easy to extend existing models with relatively few samples to new languages, it is easier to ensure desired properties for the vector space, and the hardware requirements for training is lower. We demonstrate the effectiveness of our approach for 10 languages from various language families. Code to extend sentence embeddings models to more than 400 languages is publicly available.[1]

1 Introduction

Mapping sentences or short text paragraphs to a dense vector space, such that similar sentences are close, has wide applications in NLP. It can be used for information retrieval, clustering, automatic essay scoring, and for semantic textual similarity.

However, most existent sentence embeddings models are monolingual, usually only for English, as applicable training data for other languages is scarce. For multi- and cross-lingual scenarios, only few sentence embeddings models exist.

In this publication, we present a new method that allows us to extend existent sentence embeddings models to new languages. We require a teacher model M for source language s and a set of parallel (translated) sentences $((s_1, t_1), ..., (s_n, t_n))$ with t_i the translation of s_i. Note, the t_i can be in different languages. We train a new student model $\hat{M}$ such that $\hat{M}(s_i) \approx M(s_i)$ and $\hat{M}(t_i) \approx M(s_i)$ using mean squared loss. We call this approach multilingual knowledge distillation learning, as the student $\hat{M}$ distills the knowledge of the teacher M in a multilingual setup. We demonstrate that this type of training works for various language combinations as well as for multilingual setups.

The student model $\hat{M}$ learns a multilingual sentence embedding space with two important properties: 1) Vector spaces are aligned across languages, i.e., identical sentences in different languages are mapped to the same point, 2) vector space properties in the original source language from the teacher model M are adopted and transferred to other languages.

The presented approach has various advantages compared to other training approaches for multilingual sentence embeddings. LASER (Artetxe and Schwenk, 2018) trains an encoder-decoder LSTM model using a **translation task**. The output of the encoder is used as sentence embedding. While LASER works well for identifying exact translations in different languages, it works less well for accessing the similarity of sentences that are not exact translations. When using the training method of LASER, we are not able to influence the properties of the vector space, for example, we cannot design a vector space to work well for a specific clustering task. With our approach, we can first create a vector space suited for clustering on some high-resource language, and then transfer it to other languages.

Multilingual Universal Sentence Encoder (mUSE) (Chidambaram et al., 2018; Yang et al., 2019) was trained in a multi-task setup on SNLI (Bowman et al., 2015) and on over a billion

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

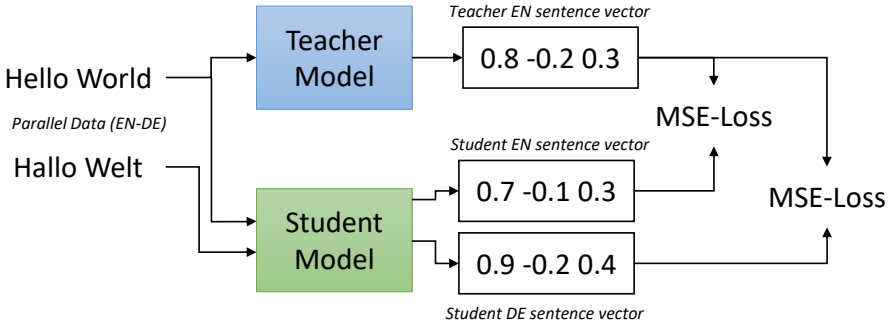

Figure 1: Given parallel data (e.g. English and German), train the student model such that the produced vectors for the English and German sentences are close to the teacher English sentence vector.

question-answer pairs from popular online forums and QA websites. In order to align the vector spaces cross-lingual, mUSE used a **translation ranking task**. Given a translation pair (s_i, t_i) and various alternative (incorrect) translations, identify the correct translation. First, multi-task learning can be difficult as it can suffer from catastrophic forgetting and balancing multiple tasks is not straight forward. Further, running the *translation ranking task* is complex and results in a huge computational overhead. Selecting random alternative translations usually leads to mediocre results. Instead, *hard negatives* (Guo et al., 2018) are required, i.e., alternative incorrect translations that have a high similarity to the correct translation. Getting these hard negative samples is non-trivial: mUSE first trained the network with random negatives samples, then used this preliminary sentence encoder to identify for each translation pair five hard negative examples (incorrect, but similar translations). It then re-trained the network. Our proposed method does not require balancing multi-task learning, nor does it require hard negative samples, making training simpler and faster.

In this publication, we use Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), which achieves state-of-the-art performance for various sentence embeddings task. SBERT is based on transformer models like BERT (Devlin et al., 2018) and applies mean pooling on the output to derive a fixed sized sentence embedding. In our experiments we use XML-RoBERTa (XLM-R) (Conneau et al., 2019), a transformer network pre-trained on 100 languages, as student model. Note, the described approach is not limited to be used with transformer models and should also work with other network architectures.

2 Training

We require a teacher model M, that maps sentences in one or more source languages s to a dense vector space. Further, we need parallel (translated) sentences $((s_1, t_1), ..., (s_n, t_n))$ with s_i a sentence in one of the source languages and t_i a sentence in one of the target languages.

We train a student model $\hat{M}$ such that $\hat{M}(s_i) \approx M(s_i)$ and $\hat{M}(t_i) \approx M(s_i)$. For a given mini-batch $\mathcal{B}$, we minimize the mean-squared loss:

$$\frac{1}{|\mathcal{B}|} \sum_{j \in \mathcal{B}} \left[(M(s_j) - \hat{M}(s_j))^2 + (M(s_j) - \hat{M}(t_j))^2 \right]$$

$\hat{M}$ could have the structure and the weights of M, or it can be a different network architecture with completely different weights. This training procedure is illustrated in Figure 1. We denote trained models with $\hat{M} \leftarrow M$, as the student model $\hat{M}$ learns the representation of the teacher model M.

In our experiments, we mainly use an English SBERT model as teacher model M and use XLM-RoBERTa (XLM-R) as student model $\hat{M}$. The English BERT models have a wordpiece vocabulary size of 30k mainly consisting of English tokens. Using the English SBERT model as initialization for $\hat{M}$ would be suboptimal, as most words in other latin-based languages would be broken down to short character sequences, and words in non-latin alphabets would be mapped to the `UNK` token. In contrast, XLM-R uses SentencePiece[2], which avoids language specific pre-processing. Further,

[2] https://github.com/google/sentencepiece

it uses a vocabulary with 250k entries from 100 different languages. This makes XLM-R much more suitable the for initialization of the multilingual student model.

3 Training Data

In this section, we evaluate the importance of training data for making the sentence embedding model multilingual. The OPUS website[3] (Tiedemann, 2012) provides parallel data for hundreds of language pairs. In our experiments, we use the following datasets:

- **GlobalVoices:** A parallel corpus of news stories from the web site Global Voices.
- **TED2020:** We crawled the translated subtitles for about 4,000 TED talks, available in over 100 languages. Resource is available in our repository.
- **NewsCommentary:** Political and economic commentary crawled from the web site Project Syndicate, provided by WMT.
- **WikiMatrix:** Mined parallel sentences from Wikipedia in different languages (Schwenk et al., 2019). We only used pairs with scores about 1.05, as pairs below this threshold were often of bad quality.
- **Tatoeba:** Tatoeba[4] is a large database of example sentences and translations to support language learning.
- **Europarl:** Parallel sentences extracted from the European Parliament website (Koehn, 2005).
- **JW300:** Mined, parallel sentences from the magazines *Awake!* and *Watchtower* (Agić and Vulić, 2019).
- **OpenSubtitles2018:** Translated movie subtitles from opensubtitles.org (Lison and Tiedemann, 2016).
- **UNPC:** Manually translated United Nations documents from 1994 - 2014 (Ziemski et al., 2016).

Getting parallel sentence data can be challenging for some low-resource language pairs. Hence, we also experiment with bilingual dictionaries:

- **MUSE:** MUSE[5] provides 110 large-scale ground-truth bilingual dictionaries created by an internal translation tool (Conneau et al., 2017b).
- **Wikititles:** We use the Wikipedia database dumps to extract the article titles from cross-language links between Wikipedia articles. For example, the page "United States" links to the German page "Vereinigte Staaten". This gives a dictionary covering a wide rang of topics.

The data set sizes for English-German (EN-DE) and English-Arabic (EN-AR) are depicted in Table 4. For training, we balance the data set sizes by drawing for a mini batch roughly the same number of samples from each data set. Data from smaller data sets is repeated.

We trained XLM-R as our student model and used SBERT fine-tuned on English NLI and STS data[6] as our teacher model. We trained for a maximum of 20 epochs with batch size 64, 10,000 warm-up steps, and a learning rate of 2e-5. As development set, we measured the MSE loss on hold-out parallel sentences.

In (Reimers and Gurevych, 2017, 2018), we showed that the random seed can have a large impact on the performances of trained models, especially for small datasets. In the following experiments, we have quite large datasets of up to several million parallel sentences and we observed rather minor differences ($\sim$ 0.3 score points) between random seeds.

4 Experiments

In this section, we conduct experiments on two tasks: Multi- and cross-lingual semantic textual similarity (STS) and bitext retrieval. STS assigns a score for a pair of sentences, while bitext retrieval identifies parallel (translated) sentences from two large monolingual corpora.

Note, evaluating the capability of different strategies to align vector spaces across languages is non-trivial. The performance for cross-lingual tasks depends on the ability to map sentences across languages to one vector space (usually the vector space for English) as well as on the properties this source vector space has. Differences

[3] http://opus.nlpl.eu/
[4] https://tatoeba.org/
[5] https://github.com/facebookresearch/MUSE
[6] bert-base-nli-stsb-mean-tokens model from our repository

in performance can then be due a better or worse alignment between the languages or due to different properties of the (source) vector space.

We evaluate the following systems:

SBERT-nli-stsb: The output of the BERT-base model is combined with mean pooling to create a fixed-sized sentence representation (Reimers and Gurevych, 2019). It was fine-tuned on the English AllNLI (SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018)) dataset and on the English training set of the STS benchmark (Cer et al., 2017) using a siamese network structure.

mBERT / XLM-R mean: Mean pooling of the outputs for the pre-trained multilingual BERT (mBERT) and XLM-R model. These models are pre-trained on multilingual data and have a multilingual vocabulary. However, no parallel data was used.

mBERT- / XLM-R-nli-stsb: We fine-tuned XLM-R and mBERT on the (English) AllNLI and the (English) training set of the STS benchmark.

LASER: LASER (Artetxe and Schwenk, 2018) uses max-pooling over the output of a stacked LSTM-encoder. The encoder was trained in a encoder-decoder setup (machine translation setup) on parallel corpora over 93 languages.

mUSE: Multilingual Universal Sentence Encoder (Chidambaram et al., 2018) uses a dual-encoder transformer architecture and was trained on mined question-answer pairs, SNLI data, translated SNLI data, and parallel corpora over 16 languages.

mBERT- / DistilmBERT- / XLM-R ← SBERT-nli-stsb: We learn mBERT, DistilmBERT, and XLM-R to imitate the output of the English SBERT-nli-stsb model.

4.1 Multilingual Semantic Textual Similarity

The goal of semantic textual similarity (STS) is to assign for a pair of sentences a score indicating their semantic similarity. For example, a score of 0 indicates *not related* and 5 indicates *semantically equivalent*.

The multilingual STS 2017 dataset (Cer et al., 2017) contains annotated pairs for EN-EN, AR-AR, ES-ES, EN-AR, EN-ES, EN-TR. We extend this dataset by translating one sentence of each pair in the EN-EN dataset to German. Further, we use Google Translate to create the datasets EN-FR, EN-IT, and EN-NL. Samples of these machine translated versions have been checked by humans fluent in that language.

We generate sentence embeddings with the described systems and compute their similarity using cosine similarity. We then compute the Spearman's rank correlation ρ between the computed score and the gold score.

We trained a single model on 10 languages[7] on all the available training datasets. Table 1 shows the performance of this model for the extended STS 2017 dataset for the same language setup, while Table 2 shows the results for the cross-lingual setup.

SBERT-nli-sts works surprisingly well for the ES-ES data. For the AR-AR data, we see a strong performance drop. This is likely because Arabic uses a non Latin alphabet, which is mapped by BERT to out-of-vocabulary tokens. The English SBERT-nli-sts does not work well for any of the cross-lingual experiments (Table 2).

Using mBERT / XLM-R out-of-the-box with mean pooling yields rather poor performances. In the same language setup (Table 1), it achieves an average correlation of only 54.0 and 42.7. For the cross-lingual setup (Table 2), the performance drops to 27.2 and 17.8. Mean pooling of out-of-the-box BERT embeddings yields unsuitable vector spaces for comparisons with cosine similarity.

Training mBERT / XLM-R on English NLI and STS data improves significantly the performance for the same language setup and achieves a performance that is on par with LASER, that was trained on cross-lingual data. However, for the cross-lingual setup (Table 2), we see a strong performance drop and more than 10 points worse results than LASER. These models can create meaningful vector spaces for sentences in the same languages. However, these vector spaces are not well-aligned across languages, as we see in Table 2.

Using our multilingual knowledge distillation approach, we observe a slight performance drop between SBERT-nli-stsb and ← SBERT-nli-stsb for the EN-EN task. However, for the ES-ES and AR-AR task, we observe a significant improvement and the model achieves a performance similar to that on the EN-EN dataset. For cross-lingual data (Table 2), we observe significantly improved performances compared to our baselines. Further, we observe a significant improvement of about 10 points in comparison to LASER. We omitted multilingual Universal Sentence Encoder (mUSE) in

[7] AR, DE, EN, ES, FR, IT, NL, RU, TR, ZH

Model	EN-EN	ES-ES	AR-AR	Avg.
SBERT-nli-stsb	83.7	71.1	49.6	68.1
mBERT mean	54.4	56.7	50.9	54.0
XLM-R mean	50.7	51.8	25.7	42.7
mBERT-nli-stsb	80.2	83.9	65.3	76.5
XLM-R-nli-stsb	78.2	83.1	64.4	75.3
LASER	77.6	79.7	68.9	75.4
mBERT ← SBERT-nli-stsb	82.5	83.0	78.8	81.4
DistilmBERT ← SBERT-nli-stsb	82.1	84.0	77.7	81.2
XLM-R ← SBERT-nli-stsb	82.5	83.5	79.9	**82.0**

Table 1: Spearman rank correlation ρ between the cosine similarity of sentence representations and the gold labels for STS 2017 dataset. Performance is reported by convention as $\rho \times 100$.

Model	EN-AR	EN-DE	EN-TR	EN-ES	EN-FR	EN-IT	EN-NL	Avg.
SBERT-nli-stsb	1.1	37.1	6.4	25.8	40.4	16.8	28.0	22.2
mBERT mean	16.7	33.9	16.0	21.5	33.0	34.0	35.6	27.2
XLM-R mean	17.4	21.3	9.2	10.9	16.6	22.9	26.0	17.8
mBERT-nli-stsb	30.9	62.2	23.9	45.4	57.8	54.3	54.1	46.9
XLM-R-nli-stsb	44.0	59.5	42.4	54.7	63.4	59.4	66.0	55.6
LASER	66.5	64.2	72.0	57.9	69.1	70.8	68.5	67.0
mBERT ← SBERT-nli-stsb	77.2	78.9	73.2	79.2	78.8	78.9	77.3	77.6
DistilmBERT ← SBERT-nli-stsb	76.1	77.7	71.8	77.6	77.4	76.5	74.7	76.0
XLM-R ← SBERT-nli-stsb	77.8	78.9	74.0	79.7	78.5	78.9	77.7	**77.9**

Table 2: Spearman rank correlation ρ between the cosine similarity of sentence representations and the gold labels for STS 2017 dataset. Performance is reported by convention as $\rho \times 100$.

this experiment, as mUSE was trained on the underlying data from STS 2017.

In our experiments, XLM-R is slightly ahead of mBERT and DistilmBERT. mBERT and DistilmBERT use different language-specific tokenization tools, making those models more difficult to be used on raw text. In contrast, XLM-R uses a SentencePiece model that can be applied directly on raw text data for all languages. Hence, in the following experiments we only report results for XLM-R.

4.2 BUCC: bitext retrieval

Given two corpora in different languages, the task is to identify sentence pairs that are translations. A straightforward approach is to take the cosine similarity of the respective sentence embeddings and to use nearest neighbor retrieval with a threshold to find translation pairs. However, it was shown that this approach has certain issues (Guo et al., 2018).

For our experiments, we use the BUCC bitext retrieval code from LASER[8]. It implements the scoring function from Artetxe and Schwenk (2019):

$$\text{score}(x, y) = \text{margin}(\cos(x, y), \sum_{z \in \text{NN}_k(x)} \frac{cos(x, z)}{2k} + \sum_{z \in \text{NN}_k(y)} \frac{cos(y, z)}{2k})$$

with x, y the two sentence embeddings and $\text{NN}_k(x)$ denoting the k nearest neighbors of x in the other language, which are retrieved using faiss[9]. As margin function, we use $\text{margin}(a, b) = a/b$.

We use the dataset from the BUCC mining task (Zweigenbaum et al., 2017, 2018), with the goal of extracting parallel sentences between an English corpus and four other languages: German, French, Russian, and Chinese. The corpora consist of 150K - 1.2M sentences for each language with about 2-3% of the sentences being parallel. The data is split into training and test sets. The training set is used to find a threshold for the score function. Pairs above the threshold are returned as parallel sentences. Performance is measured using F_1 score.

Results are shown in Table 3. Using mean pooling directly on mBERT / XLM-R produces low scores. XLM-R mean achieves only an F_1 of 11.6.

[8] https://github.com/facebookresearch/LASER/

[9] https://github.com/facebookresearch/faiss

Model	DE-EN	FR-EN	RU-EN	ZH-EN	Avg.
mBERT mean	44.1	47.2	38.0	37.4	41.7
XLM-R mean	5.2	6.6	22.1	12.4	11.6
mBERT-nli-stsb	38.9	39.5	26.4	30.2	33.7
XLM-R-nli-stsb	44.0	51.0	51.5	44.0	47.6
Knowledge Distillation					
XLM-R ← SBERT-nli-stsb	86.8	84.4	86.3	85.1	85.7
Other systems					
mUSE	88.5	86.3	89.1	86.9	87.7
LASER	95.4	92.4	92.3	91.7	92.9

Table 3: F_1 score on the BUCC bitext mining task.

While training on English NLI and STS data improves the performance for XLM-R (XLM-R-nli-stsb), it reduces the performance for mBERT. It is unclear why mBERT mean and XLM-R mean produce vastly different scores and why training on NLI data improves the cross-lingual performance for XLM, while reducing the performance for mBERT. But in conclusion, we see that mBERT / XLM-R do not have well aligned vector spaces and training only on English data is not sufficient for cross-lingual tasks.

Using our multilingual knowledge distillation method (XLM-R ← SBERT-nli-stsb), we were able to significantly improve the performance compared to the mBERT / XLM-R model trained only on English data. However, mUSE outperforms our models, and LASER significantly outperforms the mUSE models.

The imitated `SBERT-nli-stsb` model creates a vector space such that semantically similar sentences are close. However, sentences with similar meanings must not be translations of each other. For example, in the BUCC data, the following pair is not labeled as parallel text:

- Olympischen Jugend-Sommerspiele fanden vom 16. bis 28. August 2014 in Nanjing (China) statt. (en: *Summer Youth Olympic Games took place from August 16 to 28, 2014 in Nanjing (China)*)
- China hosted the 2014 Youth Olympic Games.

Both sentences are semantically similar, hence our model assigned a high similarity score. But the pair is not a translation, as some details are missing (exact dates and location).

These results stress the point that there is no single sentence vector space universally suitable for every application. LASER was trained on translation data, hence, it works well to identify perfect translations. However, it performs less well for the task of STS when it has to score sentence pairs that are only to some degree similar. In contrast, SBERT-nli-stsb works well to judge the semantic similarity of sentences, but it has difficulties to distinguish between translations and non-translation pairs with high similarities. In general, it is important to use the sentence embeddings method with the right properties for the desired downstream task.

We noticed that several positive pairs are missing in the BUCC dataset. We analyzed for SBERT, mUSE, and LASER 20 false positive DE-EN pairs each, i.e., we analyzed pairs with high similarities according to the embeddings method but which are not translations according to the dataset. For 57 out of 60 pairs, we would judge them as valid, high-quality translations. This issue comes from the way BUCC was constructed: It consists of a parallel part, drawn from the News Commentary dataset, and sentences drawn from Wikipedia, which are judged as non-parallel. However, it is not ensured that the sentences from Wikipedia are in fact non-parallel. The systems successfully returned parallel pairs from the Wikipedia part of the dataset. Results based on the BUCC dataset should be judged with care. It is unclear how many parallel sentences are in the Wikipedia part of the dataset and how this affects the scores.

5 Evaluation of Training Datasets

To evaluate the suitability of the different training sets, we trained bilingual XLM-R models for EN-DE and EN-AR on the described training datasets. English and German are fairly similar languages and have a large overlap in their alphabets, while English and Arabic are dissimilar languages with distinct alphabets. We evaluate the performance on the STS 2017 dataset.

The results for training on the full datasets

are depicted in Table 4. In Table 5, we trained the models only on the first k sentences of the TED2020 dataset.

First, we observe that the bilingual models are slightly better than the model trained for 10 languages (section 4.1): 2.2 points improvement for EN-DE and 1.2 points improvement for EN-AR. Conneau et al. (2019) calls this *curse of multilinguality*, where adding more languages to a model can degrade the performance as the capacity of the model remains the same.

Dataset	#DE	EN-DE	#AR	EN-AR
XLM-R mean	-	21.3	-	17.4
XLM-R-nli-stsb	-	59.5	-	44.0
MUSE Dict	101k	75.8	27k	68.8
Wikititles Dict	545k	71.4	748k	67.9
MUSE + Wikititles	646k	76.0	775k	69.1
GlobalVoices	37k	78.1	29k	68.6
TED2020	483k	80.4	774k	78.0
NewsCommentary	118k	77.7	7k	57.4
WikiMatrix	276k	79.4	385k	75.4
Tatoeba	303k	79.5	27k	76.7
Europarl	736k	78.7	-	-
JW300	1,399k	80.0	382k	74.0
UNPC	-	-	8M	66.1
OpenSubtitles	21M	79.8	28M	78.8
All datasets	25M	81.4	38M	79.0

Table 4: Data set sizes for the EN-DE / EN-AR sections. Performance (Spearman rank correlation) of XLM-R ← SBERT-nli-stsb on the STS 2017 dataset.

Dataset size	EN-DE	EN-AR
XLM-R mean	21.3	17.4
XLM-R-nli-stsb	59.5	44.0
1k	71.5	48.4
5k	74.5	59.6
10k	77.0	69.5
25k	80.0	70.2
Full TED2020	80.4	78.0

Table 5: Performance on STS 2017 dataset when trained with reduced dataset sizes of the TED2020 dataset.

For the similar languages EN-DE we observe only minor differences between the training datasets. It appears that the domain of the training data (news, subtitles, parliamentary debates, magazines) is not that important. As shown in Table 5, only little training data for similar languages is necessary. With only 1,000 parallel sentences, we achieve already a score of 71.8. With 25,000 sentences, we achieve a performance nearly on-par with the full German training set of 25 Million parallel sentences.

For the dissimilar languages English and Arabic, the results are less conclusive. Table 4 shows that more data does not necessarily lead to better results. With the Tatoeba dataset (only 27,000 parallel sentences), we achieve a score of 76.7, while with the UNPC dataset (over 8 Million sentences), we achieve only a score of 66.1. The domain and complexity of the parallel sentences are of higher importance for dissimilar languages. The results on the reduced TED2020 dataset (Table 5) show that the score improves slower for EN-AR than for EN-DE with more data.

Our experiments with bilingual dictionaries show that a significant improvement over an English-only baseline can be achieved. For EN-DE, about 94% and for EN-AR about 87% of the performance of the full dataset model can be achieved.

We conclude that for similar languages, like English and German, the training data is of minor importance. Already small datasets or even only bilingual dictionaries are sufficient to achieve a quite high performance. For dissimilar languages, like English and Arabic, the type of training data is of higher importance. Further, more data is necessary to achieve good results.

6 Related Work

Sentence embeddings are a well studied area with dozens of proposed methods. Skip-Thought (Kiros et al., 2015) trains an encoder-decoder architecture to predict the surrounding sentences. InferSent (Conneau et al., 2017a) uses labeled data of the Stanford Natural Language Inference dataset (Bowman et al., 2015) and the Multi-Genre NLI dataset (Williams et al., 2018) to train a siamese BiLSTM network with max-pooling over the output. Conneau et al. showed, that InferSent consistently outperforms unsupervised methods like SkipThought. Universal Sentence Encoder (Cer et al., 2018) trains a transformer network and augments unsupervised learning with training on SNLI. Hill et al. (2016) showed, that the task on which sentence embeddings are trained significantly impacts their quality. Previous work (Conneau et al., 2017a; Cer et al., 2018) found that the SNLI datasets are suitable for training sentence embeddings. Yang et al. (2018) presented a method to train on conversations from Reddit using siamese DAN and siamese transformer networks, which yielded good results on the STS benchmark dataset.

The previous methods have in common that

they were only trained on English. Multilingual representations have attracted significant attention in recent times. Most of it focuses on cross-lingual word embeddings (Ruder, 2017). A common approach is to train word embeddings for each language separately and to learn a linear transformation that maps them to shared space based on a bilingual dictionary (Artetxe et al., 2018). This mapping can also be learned without parallel data (Conneau et al., 2017b; Lample et al., 2017). Average word embeddings can further be improved by using concatenation of different power means (Rücklé et al., 2019).

A straightforward approach for creating cross-lingual sentence embeddings is to use a bag-of-words representation of cross-lingual word embeddings. However, Conneau et al. (2018) showed that this approach works poorly in practical cross-lingual transfer settings. LASER (Artetxe and Schwenk, 2018) uses a sequence-to-sequence encoder-decoder architecture (Sutskever et al., 2014) based on LSTM networks. It trains on parallel corpora akin to multilingual neural machine translation (Johnson et al., 2017). To create a fixed sized sentence representation, they apply max-pooling over the output of the encoder. LASER was trained for 93 languages on 16 NVIDIA V100 GPUs for about 5 days. In contrast, our models are trained on a single V100 GPU. The bilingual models are trained for about 4-8 hours, the multilingual model for about 2 days.

Multilingual Universal Sentence Encoder (mUSE)[10] (Chidambaram et al., 2018; Yang et al., 2019) is based on a dual-encoder architecture and uses either a CNN network or a transformer network. It was trained in a multi-task setup on SNLI (Bowman et al., 2015) and over 1 Billion crawled question-answer pairs from various communities. To make vector spaces aligned for different languages, they applied a translation ranking task: Given a sentence in the source language and a set of sentences in the target languages, identify the correct translation pair. To work well, hard negative examples (similar, but incorrect translations) must be included in the tranking task. mUSE was trained for 16 languages with 30 million steps.

In this publication, we extended Sentence-BERT (SBERT) (Reimers and Gurevych, 2019). SBERT is based on transformer models like BERT (Devlin et al., 2018) and fine-tunes those using a siamese network structure to create a sentence vector space with desired properties. By using the pre-trained weights from BERT, suitable sentence embeddings methods can be trained efficiently. Multilingual BERT (mBERT) was trained on 104 languages using Wikipedia, while XLM-R (Conneau et al., 2019) was trained on 100 languages using CommonCrawl. mBERT and XLM-R were not trained on any parallel data, hence, their vector spaces are not aligned: A sentence in different languages will be mapped to different points in vector space when these approaches are used out-of-the-box.

[10] https://tfhub.dev/google/universal-sentence-encoder

7 Conclusion

In this publication, we presented a method to make a monolingual sentence embeddings method multilingual with aligned vector spaces between the languages. This was achieved by using multilingual knowledge distillation: Given parallel data (s_i, t_i) and a teacher model M, we train a student model $\hat{M}$ such that $\hat{M}(s_i) \approx M(s_i)$ and $\hat{M}(t_i) \approx M(s_i)$.

We demonstrated that this approach successfully transfers properties from the source language vector space (in our case English) to various target languages. Models can be extended to multiple languages in the same training process. The approach can also be applied to multilingual teacher models M to extend those to further languages.

This stepwise training approach has the advantage that an embedding model with desired properties, for example for clustering, can first be created for a high-resource language. Then, in an independent step, it can be extended to support further languages. This decoupling significantly simplifies the training procedure compared to previous approaches.

Acknowledgments

This work has been supported by the German Research Foundation through the German-Israeli Project Cooperation (DIP, grant DA 1600/1-1 and grant GU 798/17-1).

References

Željko Agić and Ivan Vulić. 2019. JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. *AAAI Conference on Artificial Intelligence*.

Mikel Artetxe and Holger Schwenk. 2018. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *arXiv preprint arXiv:1812.10464*, abs/1812.10464.

Mikel Artetxe and Holger Schwenk. 2019. 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.

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*, pages 632–642, Lisbon, Portugal. 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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *arXiv preprint arXiv:1803.11175*.

Muthuraman Chidambaram, Yinfei Yang, Daniel Cer, Steve Yuan, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning Cross-Lingual Sentence Representations via a Multi-task Dual-Encoder Model. *arXiv preprint arXiv:1810.12836*, abs/1810.12836.

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

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017a. 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, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017b. Word Translation Without Parallel Data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel 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*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

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, Brussels, Belgium. Association for Computational Linguistics.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning Distributed Representations of Sentences from Unlabelled Data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-Thought Vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised Machine Translation Using Monolingual Corpora Only. *arXiv preprint arXiv:1711.00043*.

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 2016)*, Paris, France. European Language Resources Association (ELRA).

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2018. Why Comparing Single Performance Scores Does Not Allow to Draw Conclusions About Machine Learning Approaches. *arXiv preprint arXiv:1803.09578*, abs/1803.09578.

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.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2019. Concatenated p-mean Word Embeddings as Universal Cross-Lingual Sentence Representations. *arXiv preprint arXiv:1803.01400*.

Sebastian Ruder. 2017. A survey of cross-lingual embedding models. *arXiv preprint arXiv:1706.04902*, abs/1706.04902.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia. *arXiv preprint arXiv:11907.05791*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Jörg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

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 Hernández Ábrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Multilingual Universal Sentence Encoder for Semantic Retrieval. *arXiv preprint arXiv:1907.04307*, abs/1907.04307.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-Yi 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.

Micha 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 2016)*, Paris, France. European Language Resources Association (ELRA).

Pierre Zweigenbaum, Serge Sharoff, and Reinhar 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).

Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. Overview of the Second BUCC Shared Task: Spotting Parallel Sentences in Comparable Corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.