# Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond

**Mikel Artetxe**
University of the Basque Country
(UPV/EHU)*
mikel.artetxe@ehu.eus

**Holger Schwenk**
Facebook AI Research
schwenk@fb.com

## Abstract

We introduce an architecture to learn joint multilingual sentence representations for 93 languages, belonging to more than 30 different families and written in 28 different scripts. Our system uses a single BiLSTM encoder with a shared byte-pair encoding vocabulary for all languages, which is coupled with an auxiliary decoder and trained on publicly available parallel corpora. This enables us to learn a classifier on top of the resulting embeddings using English annotated data only, and transfer it to any of the 93 languages without any modification. Our experiments in cross-lingual natural language inference (XNLI data set), cross-lingual document classification (MLDoc data set), and parallel corpus mining (BUCC data set) show the effectiveness of our approach. We also introduce a new test set of aligned sentences in 112 languages, and show that our sentence embeddings obtain strong results in multilingual similarity search even for low-resource languages. Our implementation, the pre-trained encoder, and the multilingual test set are available at `https://github.com/facebookresearch/LASER`.

## 1 Introduction

While the recent advent of deep learning has led to impressive progress in natural language processing (NLP), these techniques are known to be particularly data-hungry, limiting their applicability in many practical scenarios. An increasingly popular approach to alleviate this issue is to first learn general language representations on unlabeled data, which are then integrated in task-specific downstream systems. This approach was first popularized by word embeddings (Mikolov et al., 2013b; Pennington et al., 2014), but has recently been superseded by sentence-level representations (Peters et al., 2018; Devlin et al., 2019). Nevertheless, all these works learn a separate model for each language and are thus unable to leverage information across different languages, greatly limiting their potential performance for low-resource languages.

In this work, we are interested in **universal language agnostic sentence embeddings**, that is, vector representations of sentences that are general with respect to two dimensions: the input language and the NLP task. The motivations for such representations are multiple: the hope that languages with limited resources benefit from joint training over many languages, the desire to perform zero-shot transfer of an NLP model from one language (typically English) to another, and the possibility to handle code-switching. To that end, we train a single encoder to handle multiple languages, so that semantically similar sentences in different languages are close in the embedding space.

Whereas previous work in multilingual NLP has been limited to either a few languages (Schwenk and Douze, 2017; Yu et al., 2018) or specific applications like typology prediction (Malaviya et al., 2017) or machine translation (Neubig and Hu, 2018), we learn general purpose sentence representations for 93 languages (see Table 1). Using a single pre-trained BiLSTM encoder for all 93 languages, we obtain very strong results in various scenarios without any fine-tuning, including cross-lingual natural language inference (XNLI data set), cross-lingual classification (MLDoc data set), bitext mining (BUCC data set), and a new multilingual similarity search data set we introduce covering 112 languages. To the best

*This work was performed during an internship at Facebook AI Research.

Transactions of the Association for Computational Linguistics, vol. 7, pp. 597–610, 2019. https://doi.org/10.1162/tacl_a_00288.
Action Editor: James Henderson. Submission batch: 4/2019; Revision batch: 7/2019; Published 9/2019.

| | **af** | **am** | **ar** | **ay** | **az** | **be** | **ber** | **bg** | **bn** | **br** | **bs** | **ca** | **cbk** | **cs** | **da** | **de** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **train sent.** | 67k | 88k | 8.2M | 14k | 254k | 5k | 62k | 4.9M | 913k | 29k | 4.2M | 813k | 1k | 5.5M | 7.9M | 8.7M |
| **en→xx err.** | 11.20 | 60.71 | 8.30 | n/a | 44.10 | 31.20 | 29.80 | 4.50 | 10.80 | 83.50 | 3.95 | 4.00 | 24.20 | 3.10 | 3.90 | 0.90 |
| **xx→en err.** | 9.90 | 55.36 | 7.80 | n/a | 23.90 | 36.50 | 33.70 | 5.40 | 10.00 | 84.90 | 3.11 | 4.20 | 21.70 | 3.80 | 4.00 | 1.00 |
| **test sent.** | 1000 | 168 | 1000 | – | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 354 | 1000 | 1000 | 1000 | 1000 | 1000 |
| | **dtp** | **dv** | **el** | **en** | **eo** | **es** | **et** | **eu** | **fi** | **fr** | **ga** | **gl** | **ha** | **he** | **hi** | **hr** |
| **train sent.** | 1k | 90k | 6.5M | 2.6M | 397k | 4.8M | 5.3M | 1.2M | 7.9M | 8.8M | 732 | 349k | 127k | 4.1M | 288k | 4.0M |
| **en→xx err.** | 92.10 | n/a | 5.30 | n/a | 2.70 | 1.90 | 3.20 | 5.70 | 3.70 | 4.40 | 93.80 | 4.60 | n/a | 8.10 | 5.80 | 2.80 |
| **xx→en err.** | 93.50 | n/a | 4.80 | n/a | 2.80 | 2.10 | 3.40 | 5.00 | 3.70 | 4.30 | 95.80 | 4.40 | n/a | 7.60 | 4.80 | 2.70 |
| **test sent.** | 1000 | – | 1000 | – | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | – | 1000 | 1000 | 1000 |
| | **hu** | **hy** | **ia** | **id** | **ie** | **io** | **is** | **it** | **ja** | **ka** | **kab** | **kk** | **km** | **ko** | **ku** | **kw** |
| **train sent.** | 5.3M | 6k | 9k | 4.3M | 3k | 3k | 2.0M | 8.3M | 3.2M | 296k | 15k | 4k | 625 | 1.4M | 50k | 2k |
| **en→xx err.** | 3.90 | 59.97 | 5.40 | 5.20 | 14.70 | 17.40 | 4.40 | 4.60 | 3.90 | 60.32 | 39.10 | 80.17 | 77.01 | 10.60 | 80.24 | 91.90 |
| **xx→en err.** | 4.00 | 67.79 | 4.10 | 5.80 | 12.80 | 15.20 | 4.40 | 4.80 | 5.40 | 67.83 | 44.70 | 82.61 | 81.72 | 11.50 | 85.37 | 93.20 |
| **test sent.** | 1000 | 742 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 746 | 1000 | 575 | 722 | 1000 | 410 | 1000 |
| | **kzj** | **la** | **lfn** | **lt** | **lv** | **mg** | **mhr** | **mk** | **ml** | **mr** | **ms** | **my** | **nb** | **nds** | **nl** | **oc** |
| **train sent.** | 560 | 19k | 2k | 3.2M | 2.0M | 355k | 1k | 4.2M | 373k | 31k | 2.9M | 2k | 4.1M | 12k | 8.4M | 3k |
| **en→xx err.** | 91.60 | 41.60 | 35.90 | 4.10 | 4.50 | n/a | 87.70 | 5.20 | 3.35 | 9.00 | 3.40 | n/a | 1.30 | 18.60 | 3.10 | 39.20 |
| **xx→en err.** | 94.10 | 41.50 | 35.10 | 3.40 | 4.70 | n/a | 91.50 | 5.40 | 2.91 | 8.00 | 3.80 | n/a | 1.10 | 15.60 | 4.30 | 38.40 |
| **test sent.** | 1000 | 1000 | 1000 | 1000 | 1000 | – | 1000 | 1000 | 687 | 1000 | 1000 | – | 1000 | 1000 | 1000 | 1000 |
| | **pl** | **ps** | **pt** | **ro** | **ru** | **sd** | **si** | **sk** | **sl** | **so** | **sq** | **sr** | **sv** | **sw** | **ta** | **te** |
| **train sent.** | 5.5M | 4.9M | 8.3M | 4.9M | 9.3M | 91k | 796k | 5.2M | 5.2M | 85k | 3.2M | 4.0M | 7.8M | 173k | 42k | 33k |
| **en→xx err.** | 2.00 | 7.20 | 4.70 | 2.50 | 4.90 | n/a | n/a | 3.10 | 4.50 | n/a | 1.80 | 4.30 | 3.60 | 45.64 | 31.60 | 18.38 |
| **xx→en err.** | 2.40 | 6.00 | 4.90 | 2.70 | 5.90 | n/a | n/a | 3.70 | 3.77 | n/a | 2.30 | 5.00 | 3.20 | 39.23 | 29.64 | 22.22 |
| **test sent.** | 1000 | 1000 | 1000 | 1000 | 1000 | – | – | 1000 | 823 | – | 1000 | 1000 | 1000 | 390 | 307 | 234 |
| | **tg** | **th** | **tl** | **tr** | **tt** | **ug** | **uk** | **ur** | **uz** | **vi** | **wuu** | **yue** | **zh** | | | |
| **train sent.** | 124k | 4.1M | 36k | 5.7M | 119k | 88k | 1.4M | 746k | 118k | 4.0M | 2k | 4k | 8.3M | | | |
| **en→xx err.** | n/a | 4.93 | 47.40 | 2.30 | 72.00 | 59.90 | 5.80 | 20.00 | 82.24 | 3.40 | 25.80 | 37.00 | 4.10 | | | |
| **xx→en err.** | n/a | 4.20 | 51.50 | 2.60 | 65.70 | 49.60 | 5.10 | 16.20 | 80.37 | 3.00 | 25.20 | 38.90 | 5.00 | | | |
| **test sent.** | – | 548 | 1000 | 1000 | 1000 | 1000 | 1000 | 1000 | 428 | 1000 | 1000 | 1000 | 1000 | | | |

Table 1: List of the 93 languages along with their training size, the resulting similarity error rate on Tatoeba, and the number of sentences in it. Dashes denote language pairs excluded for containing fewer than 100 test sentences.

of our knowledge, this is the first exploration of general purpose massively multilingual sentence representations across a large variety of tasks.

## 2 Related Work

Following the success of word embeddings (Mikolov et al., 2013b; Pennington et al., 2014), there has been an increasing interest in learning continuous vector representations of longer linguistic units like sentences (Le and Mikolov, 2014; Kiros et al., 2015). These sentence embeddings are commonly obtained using a recurrent neural network (RNN) encoder, which is typically trained in an unsupervised way over large collections of unlabeled corpora. For instance, the skip-thought model of Kiros et al. (2015) couples the encoder with an auxiliary decoder, and trains the entire system to predict the surrounding sentences over a collection of books. It was later shown that more competitive results could be obtained by training the encoder over labeled natural language inference (NLI) data (Conneau et al., 2017). This was later extended to multitask learning, combining different training objectives like that of skip-thought, NLI, and machine translation (Cer et al., 2018; Subramanian et al., 2018).

While the previous methods consider a single language at a time, multilingual representations have recently attracted a large attention. Most of this research focuses on cross-lingual word embeddings (Ruder et al., 2017), which are commonly learned jointly from parallel corpora (Gouws et al., 2015; Luong et al., 2015). An alternative approach that is becoming increasingly popular is to separately train word embeddings for each language, and map them to a shared space based on a bilingual dictionary (Mikolov et al., 2013a; Artetxe et al., 2018a) or even in a fully unsupervised manner (Conneau et al., 2018a; Artetxe et al., 2018b). Cross-lingual word embeddings are often used to build bag-of-word representations of longer linguistic units by taking their respective (IDF-weighted) average (Klementiev et al., 2012; Dufter et al., 2018). Although this approach has the advantage of requiring weak or no cross-lingual signal, it has been shown that the resulting sentence embeddings work poorly in practical cross-lingual transfer settings (Conneau et al., 2018b).

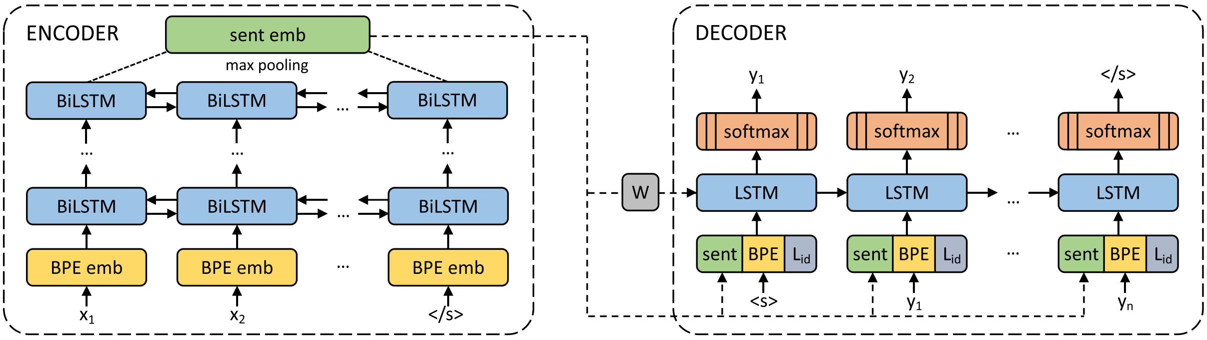

Figure 1: Architecture of our system to learn multilingual sentence embeddings.

A more competitive approach that we follow here is to use a sequence-to-sequence encoder-decoder architecture (Schwenk and Douze, 2017; Hassan et al., 2018). The full system is trained end-to-end on parallel corpora akin to multilingual neural machine translation (Johnson et al., 2017): The encoder maps the source sequence into a fixed-length vector representation, which is used by the decoder to create the target sequence. This decoder is then discarded, and the encoder is kept to embed sentences in any of the training languages. While some proposals use a separate encoder for each language (Schwenk and Douze, 2017), sharing a single encoder for all languages also gives strong results (Schwenk, 2018).

Nevertheless, most existing work is either limited to a few, rather close languages (Schwenk and Douze, 2017; Yu et al., 2018) or, more commonly, consider pairwise joint embeddings with English and one foreign language (España-Bonet et al., 2017; Guo et al., 2018). To the best of our knowledge, existing work on learning multilingual representations for a large number of languages is limited to word embeddings (Ammar et al., 2016; Dufter et al., 2018) specific applications like typology prediction (Malaviya et al., 2017) or machine translation (Neubig and Hu, 2018)—ours being the first paper exploring general purpose massively multilingual sentence representations.

All the previous approaches learn a fixed-length representation for each sentence. A recent research line has obtained very strong results using variable-length representations instead, consisting of contextualized embeddings of the words in the sentence (Dai and Le, 2015; Peters et al., 2018; Howard and Ruder, 2018; Devlin et al., 2019). For that purpose, these methods train either an RNN or self-attentional encoder over unnanotated corpora using some form of language modeling. A classifier can then be learned on top of the resulting encoder, which is commonly further fine-tuned during this supervised training. Concurrent to our work, Lample and Conneau (2019) propose a cross-lingual extension of these models, and report strong results in cross-lingual NLI, machine translation, and language modeling. In contrast, our focus is on scaling to a large number of languages, for which we argue that fixed-length approaches provide a more versatile and compatible representation form.[1] Also, our approach achieves strong results without task-specific fine-tuning, which makes it interesting for tasks with limited resources.

## 3 Proposed Method

We use a single, language-agnostic BiLSTM encoder to build our sentence embeddings, which is coupled with an auxiliary decoder and trained on parallel corpora. In Sections 3.1 to 3.3, we describe its architecture, our training strategy to scale to 93 languages, and the training data used for that purpose.

### 3.1 Architecture

Figure 1 illustrates the architecture of the proposed system, which is based on Schwenk (2018). As it can be seen, sentence embeddings are obtained by applying a max-pooling operation over the output of a BiLSTM encoder. These sentence embeddings are used to initialize the decoder LSTM through a linear transformation, and are also concatenated to its input embeddings at every time step. Note that there is no other connection

[1] For instance, there is not always a one-to-one correspondence among words in different languages (e.g., a single word of a morphologically complex language might correspond to several words of a morphologically simple language), so having a separate vector for each word might not transfer as well across languages.

between the encoder and the decoder, as we want all relevant information of the input sequence to be captured by the sentence embedding.

We use a single encoder and decoder in our system, which are shared by all languages involved. For that purpose, we build a joint byte-pair encoding (BPE) vocabulary with 50k operations, which is learned on the concatenation of all training corpora. This way, the encoder has no explicit signal on what the input language is, encouraging it to learn language independent representations. In contrast, the decoder takes a language ID embedding that specifies the language to generate, which is concatenated to the input and sentence embeddings at every time step.

Scaling up to almost one hundred languages calls for an encoder with sufficient capacity. In this paper, we limit our study to a stacked BiLSTM with 1 to 5 layers, each 512-dimensional. The resulting sentence representations (after concatenating both directions) are 1024-dimensional. The decoder has always one layer of dimension 2048. The input embedding size is set to 320, and the language ID embedding has 32 dimensions.

### 3.2 Training Strategy

In preceding work (Schwenk and Douze, 2017; Schwenk, 2018), each input sentence was jointly translated into all other languages. However, this approach has two obvious drawbacks when trying to scale to a large number of languages. First, it requires an N-way parallel corpus, which is difficult to obtain for all languages. Second, it has a quadratic cost with respect to the number of languages, making training prohibitively slow as the number of languages is increased. In our preliminary experiments, we observed that similar results can be obtained using only two target languages.[2] At the same time, we relax the requirement for N-way parallel corpora by considering separate alignments for each language combination.

Training minimizes the cross-entropy loss on the training corpus, alternating over all combinations of the languages involved. For that purpose, we use Adam with a constant learning rate of 0.001 and dropout set to 0.1, and train for a fixed number of epochs. Our implementation is based on `fairseq`,[3] and we make use of its multi-GPU support to train on 16 NVIDIA V100 GPUs with a total batch size of 128,000 tokens. Unless otherwise specified, we train our model for 17 epochs, which takes about 5 days. Stopping training earlier decreases the overall performance only slightly.

### 3.3 Training Data and Pre-processing

As described in Section 3.2, training requires bitexts aligned with two target languages. We choose English and Spanish for that purpose, as most of the data are aligned with these languages.[4] We collect training corpora for 93 input languages by combining the Europarl, United Nations, OpenSubtitles2018, Global Voices, Tanzil, and Tatoeba corpuses, which are all publicly available on the OPUS Web site[5] (Tiedemann, 2012). Appendix A provides a more detailed description of this training data, and Table 1 summarizes the list of all languages covered and the size of the bitexts. Our training data comprises a total of 223 million parallel sentences. All pre-processing is done with Moses tools:[6] punctuation normalization, removing non-printing characters, and tokenization. As the only exception, Chinese and Japanese were segmented with Jieba[7] and Mecab,[8] respectively. All the languages are kept in their original script with the exception of Greek, which we romanize into the Latin alphabet. It is important to note that the joint encoder itself has no information on the language or writing script of the tokenized input texts. It is even possible to mix multiple languages in one sentence.

## 4 Experimental Evaluation

In contrast with the well-established evaluation frameworks for English sentence representations (Conneau et al., 2017; Wang et al., 2018), there

[2] Note that, if we had a single target language, the only way to train the encoder for that language would be auto-encoding, which we observe to work poorly. Having two target languages avoids this problem.

[3] https://github.com/pytorch/fairseq

[4] Note that it is not necessary that all input languages are systematically aligned with both target languages. Once we have several languages with both alignments, the joint embedding is well conditioned, and we can add more languages with one alignment only, usually English.

[5] http://opus.nlpl.eu

[6] http://www.statmt.org/moses

[7] https://github.com/fxsjy/jieba

[8] https://github.com/taku910/mecab

| | | EN | EN → XX | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | fr | es | de | el | bg | ru | tr | ar | vi | th | zh | hi | sw | ur |
| **Zero-Shot Transfer, one NLI system for all languages:** | | | | | | | | | | | | | | | | |
| Conneau et al. | X-BiLSTM | 73.7 | 67.7 | 68.7 | 67.7 | 68.9 | 67.9 | 65.4 | 64.2 | 64.8 | 66.4 | 64.1 | 65.8 | 64.1 | 55.7 | 58.4 |
| (2018b) | X-CBOW | 64.5 | 60.3 | 60.7 | 61.0 | 60.5 | 60.4 | 57.8 | 58.7 | 57.5 | 58.8 | 56.9 | 58.8 | 56.3 | 50.4 | 52.2 |
| BERT uncased* | Transformer | 81.4 | – | 74.3 | 70.5 | – | – | – | – | 62.1 | – | – | 63.8 | – | – | 58.3 |
| Proposed method | BiLSTM | 73.9 | **71.9** | 72.9 | 72.6 | **72.8** | **74.2** | **72.1** | **69.7** | **71.4** | **72.0** | **69.2** | 71.4 | **65.5** | **62.2** | 61.0 |
| **Translate test, one English NLI system:** | | | | | | | | | | | | | | | | |
| Conneau et al. (2018b) | BiLSTM | 73.7 | 70.4 | 70.7 | 68.7 | 69.1 | 70.4 | 67.8 | 66.3 | 66.8 | 66.5 | 64.4 | 68.3 | 64.2 | 61.8 | 59.3 |
| BERT uncased* | Transformer | 81.4 | – | 74.9 | 74.4 | – | – | – | – | 70.4 | – | – | 70.1 | – | – | **62.1** |
| **Translate train, separate NLI systems for each language:** | | | | | | | | | | | | | | | | |
| Conneau et al. (2018b) | BiLSTM | 73.7 | 68.3 | 68.8 | 66.5 | 66.4 | 67.4 | 66.5 | 64.5 | 65.8 | 66.0 | 62.8 | 67.0 | 62.1 | 58.2 | 56.6 |
| BERT cased* | Transformer | **81.9** | – | **77.8** | **75.9** | – | – | – | – | 70.7 | – | *68.9*† | **76.6** | – | – | 61.6 |

Table 2: Test accuracies on the XNLI cross-lingual natural language inference data set. All results from Conneau et al. (2018b) correspond to max-pooling, which outperforms the last-state variant in all cases. Results involving machine translation do not use a multilingual model and are not directly comparable with zero-shot transfer. Overall best results are in bold, the best ones in each group are underlined.
* Results for BERT (Devlin et al., 2019) are extracted from its GitHub README.[9]
† Monolingual BERT model for Thai from `https://github.com/ThAIKeras/bert`.

is not yet a commonly accepted standard to evaluate multilingual sentence embeddings. The most notable effort in this regard is arguably the XNLI data set (Conneau et al., 2018b), which evaluates the transfer performance of an NLI model trained on English data over 14 additional test languages (Section 4.1). So as to obtain a more complete picture, we also evaluate our embeddings in cross-lingual document classification (MLDoc, Section 4.2), and bitext mining (BUCC, Section 4.3). However, all these data sets only cover a subset of our 93 languages, so we also introduce a new test set for multilingual similarity search in 112 languages, including several languages for which we have no training data but whose language family is covered (Section 4.4). We remark that we use the same pre-trained BiLSTM encoder for all tasks and languages without any fine-tuning.

### 4.1 XNLI: Cross-lingual NLI

NLI has become a widely used task to evaluate sentence representations (Bowman et al., 2015; Williams et al., 2018). Given two sentences, a premise and a hypothesis, the task consists in deciding whether there is an *entailment*, *contradiction*, or *neutral* relationship between them. XNLI is a recent effort to create a data set similar to the English MultiNLI for several languages (Conneau et al., 2018b). It consists of 2,500 development and 5,000 test instances translated from English into 14 languages by professional translators, making results across different languages directly comparable.

We train a classifier on top of our multilingual encoder using the usual combination of the two sentence embeddings: $(p, h, p \cdot h, |p - h|)$, where $p$ and $h$ are the premise and hypothesis. For that purpose, we use a feed-forward neural network with two hidden layers of size 512 and 384, trained with Adam. All hyperparameters were optimized on the English XNLI development corpus only, and then the same classifier was applied to all languages of the XNLI test set. As such, we did not use any training or development data in any of the foreign languages. Note, moreover, that the multilingual sentence embeddings are fixed and not fine-tuned on the task or the language.

We report our results in Table 2, along with several baselines from Conneau et al. (2018b) and the multilingual BERT model (Devlin et al., 2019).[9] Our proposed method obtains the best results in zero-shot cross-lingual transfer for all languages but Spanish. Moreover, our transfer results are strong and homogeneous across all languages: For 11 of them, the zero-short performance

[9] Note that the multilingual variant of BERT is not discussed in its paper (Devlin et al., 2019). Instead, the reported results were extracted from the README of the official GitHub project at `https://github.com/google-research/bert/blob/master/multilingual.md` on July 5, 2019.

| | | EN | EN → XX | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | de | es | fr | it | ja | ru | zh |
| Schwenk | MultiCCA + CNN | **92.20** | 81.20 | 72.50 | 72.38 | 69.38 | **67.63** | 60.80 | **74.73** |
| and Li | BiLSTM (Europarl) | 88.40 | 71.83 | 66.65 | 72.83 | 60.73 | - | - | - |
| (2018) | BiLSTM (UN) | 88.83 | - | 69.50 | 74.52 | - | - | 61.42 | 71.97 |
| Proposed method | | 89.93 | **84.78** | **77.33** | **77.95** | **69.43** | 60.30 | **67.78** | 71.93 |

Table 3: Accuracies on the MLDoc zero-shot cross-lingual document classification task (test set).

is (at most) 5% lower than the one on English, including distant languages like Arabic, Chinese, and Vietnamese, and we also achieve remarkable good results on low-resource languages like Swahili. In contrast, BERT achieves excellent results on English, outperforming our system by 7.5 points, but its transfer performance is much weaker. For instance, the loss in accuracy for both Arabic and Chinese is 2.5 points for our system, compared with 19.3 and 17.6 points for BERT.[10] Finally, we also outperform all baselines of Conneau et al. (2018b) by a substantial margin, with the additional advantage that we use a single pre-trained encoder, whereas X-BiLSTM learns a separate encoder for each language.

We also provide results involving Machine Translation (MT) from Conneau et al. (2018b). This can be done in two ways: 1) translate the test data into English and apply the English NLI classifier, or 2) translate the English training data and train a separate NLI classifier for each language. Note that we are not evaluating multilingual sentence embeddings anymore, but rather the quality of the MT system and a monolingual model. Moreover, the use of MT incurs an important overhead with either strategy: Translating test makes inference substantially more expensive, whereas translating train results in a separate model for each language. As shown in Table 2, our approach outperforms all translation baselines of Conneau et al. (2018b). We also outperform MT BERT for Arabic and Thai, and are very close for Urdu. Thanks to its multilingual nature, our system can also handle premises and hypothesis in different languages. As reported in Appendix B, the proposed method obtains very strong results in these settings, even for distant language combinations like French–Chinese.

## 4.2 MLDoc: Cross-lingual Classification

Cross-lingual document classification is a typical application of multilingual representations. In order to evaluate our sentence embeddings in this task, we use the MLDoc data set of Schwenk and Li (2018), which is an improved version of the Reuters benchmark (Lewis et al., 2004; Klementiev et al., 2012) with uniform class priors and a wider language coverage. There are 1,000 training and development documents and 4,000 test documents for each language, divided in 4 different genres. Just as with the XNLI evaluation, we consider the zero-shot transfer scenario: We train a classifier on top of our multilingual encoder using the English training data, optimizing hyper-parameters on the English development set, and evaluating the resulting system in the remaining languages. We use a feed-forward neural network with one hidden layer of 10 units.

As shown in Table 3, our system obtains the best published results for 5 of the 7 transfer languages. We believe that our weaker performance on Japanese can be attributed to the domain and sentence length mismatch between MLDoc and the parallel corpus we use for this language.

## 4.3 BUCC: Bitext Mining

Bitext mining is another natural application for multilingual sentence embeddings. Given two comparable corpora in different languages, the task consists of identifying sentence pairs that are translations of each other. For that purpose, one would commonly score sentence pairs by taking the cosine similarity of their respective embeddings, so parallel sentences can be extracted through nearest neighbor retrieval and filtered by

[10] Concurrent to our work, Lample and Conneau (2019) report superior results using another variant of BERT, outperforming our method by 4.5 points in average. However, note that these results are not fully comparable because 1) their system uses development data in the foreign languages, whereas our approach is fully zero-shot, 2) their approach requires fine-tuning on the task, 3) our system handles a much larger number of languages, and 4) our transfer performance is substantially better (an average loss of 4 vs 10.6 points with respect to the respective English system).

|  | TRAIN |  |  |  | TEST |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | de-en | fr-en | ru-en | zh-en | de-en | fr-en | ru-en | zh-en |
| Azpeitia et al. (2017) | 83.33 | 78.83 | - | - | 83.74 | 79.46 | - | - |
| Grégoire and Langlais (2017) | - | 20.67 | - | - | - | 20 | - | - |
| Zhang and Zweigenbaum (2017) | - | - | - | 43.48 | - | - | - | 45.13 |
| Azpeitia et al. (2018) | 84.27 | 80.63 | 80.89 | 76.45 | 85.52 | 81.47 | 81.30 | 77.45 |
| Bouamor and Sajjad (2018) | - | 75.2 | - | - | - | 76.0 | - | - |
| Chongman Leong and Chao (2018) | - | - | - | 58.54 | - | - | - | 56 |
| Schwenk (2018) | 76.1 | 74.9 | 73.3 | 71.6 | 76.9 | 75.8 | 73.8 | 71.6 |
| Artetxe and Schwenk (2018) | 94.84 | 91.85 | 90.92 | 91.04 | 95.58 | 92.89 | 92.03 | **92.57** |
| Proposed method | **95.43** | **92.40** | **92.29** | **91.20** | **96.19** | **93.91** | **93.30** | 92.27 |

Table 4: F1 scores on the BUCC mining task.

setting a fixed threshold over this score (Schwenk, 2018). However, it was recently shown that this approach suffers from scale inconsistency issues (Guo et al., 2018), and Artetxe and Schwenk (2018) proposed the following alternative score addressing it:

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

where $x$ and $y$ are the source and target sentences, and $\text{NN}_k(x)$ denotes the $k$ nearest neighbors of $x$ in the other language. The paper explores different margin functions, with *ratio* ($\text{margin}(a, b) = \frac{a}{b}$) yielding the best results. This notion of margin is related to CSLS (Conneau et al., 2018a).

We use this method to evaluate our sentence embeddings on the BUCC mining task (Zweigenbaum et al., 2017, 2018), using exact same hyper-parameters as Artetxe and Schwenk (2018). The task consists in extracting parallel sentences from a comparable corpus between English and four foreign languages: German, French, Russian, and Chinese. The data set consists of 150 K to 1.2 M sentences for each language, split into a sample, training and test set, with about 2–3% of the sentences being parallel. As shown in Table 4, our system establishes a new state-of-the-art for all language pairs with the exception of English-Chinese test. We also outperform Artetxe and Schwenk (2018) themselves, who use two separate models covering 4 languages each. Not only are our results better, but our model also covers many more languages, so it can potentially be used to mine bitext for any combination of the 93 languages supported.

### 4.4 Tatoeba: Similarity Search

Although XNLI, MLDoc, and BUCC are well-established benchmarks with comparative results available, they only cover a small subset of our 93 languages. So as to better assess the performance of our model in all these languages, we introduce a new test set of similarity search for 112 languages based on the Tatoeba corpus. The data set consists of up to 1,000 English-aligned sentence pairs for each language. Appendix C describes how the data set was constructed in more details. Evaluation is done by finding the nearest neighbor for each sentence in the other language according to cosine similarity and computing the error rate.

We report our results in Table 1. Contrasting these results with those of XNLI, one would assume that similarity error rates below 5% are indicative of strong downstream performance.[11] This is the case for 37 languages, there are 48 languages with an error rate below 10% and 55 with less than 20%. There are only 15 languages with error rates above 50%. Additional result analysis is given in Appendix D.

We believe that our competitive results for many low-resource languages are indicative of the benefits of joint training, which is also supported by our ablation results in Section 5.3. In relation to that, Appendix E reports similarity search results for 29 additional languages without any training data, showing that our encoder can also generalize to unseen languages to some extent as long as it was trained on related languages.

## 5 Ablation Experiments

In this section, we explore different variants of our approach and study the impact on the performance

[11] We consider the average of en→xx and xx→en

| Depth | Tatoeba<br>Err [%] | BUCC<br>F1 | MLDoc<br>Acc [%] | XNLI-en<br>Acc [%] | XNLI-xx<br>Acc [%] |
|---|---|---|---|---|---|
| 1 | 37.96 | 89.95 | 69.42 | 70.94 | 64.54 |
| 3 | 28.95 | 92.28 | 71.64 | 72.83 | 68.43 |
| 5 | **26.31** | **92.83** | **72.79** | **73.67** | **69.92** |

Table 5: Impact of the depth of the BiLSTM encoder.

| NLI<br>obj. | Tatoeba<br>Err [%] | BUCC<br>F1 | MLDoc<br>Acc [%] | XNLI-en<br>Acc [%] | XNLI-xx<br>Acc [%] |
|---|---|---|---|---|---|
| – | **26.31** | 92.83 | 72.79 | 73.67 | **69.92** |
| ×1 | 26.89 | 93.01 | **74.51** | 73.71 | 69.10 |
| ×2 | 28.52 | **93.06** | 71.90 | 74.65 | 67.75 |
| ×3 | 27.83 | 92.98 | 73.11 | **75.23** | 61.86 |

Table 6: Multitask training with an NLI objective and different weightings.

| #langs | WMT<br>Err [%] | BUCC<br>F1 | MLDoc<br>Acc [%] | XNLI-en<br>Acc [%] | XNLI-xx<br>Acc [%] |
|---|---|---|---|---|---|
| All (93) | **0.54** | 92.83 | 72.79 | **73.67** | **69.92** |
| Eval (18) | 0.59 | **92.91** | **75.63** | 72.99 | 68.84 |

Table 7: Comparison between training on 93 languages and training on the 18 evaluation languages only.

for all our evaluation tasks. We report average results across all languages. For XNLI, we also report the accuracy on English.

### 5.1 Encoder Depth

Table 5 reports the performance on the different tasks for encoders with 1, 3, or 5 layers. We were not able to achieve good convergence with deeper models. It can be seen that all tasks benefit from deeper models, in particular XNLI and Tatoeba, suggesting that a single-layer BiLSTM has not enough capacity to encode so many languages.

### 5.2 Multitask Learning

Multitask learning has been shown to be helpful to learn English sentence embeddings (Subramanian et al., 2018; Cer et al., 2018). The most important task in this approach is arguably NLI, so we explored adding an additional NLI objective to our system with different weighting schemes. As shown in Table 6, the NLI objective leads to a better performance on the English NLI test set, but this comes at the cost of a worse cross-lingual transfer performance in XNLI and Tatoeba. The effect in BUCC is negligible.

### 5.3 Number of Training Languages

So as to better understand how our architecture scales to a large amount of languages, we train a separate model on a subset of 18 evaluation languages, and compare it to our main model trained on 93 languages. We replaced the Tatoeba corpus with the WMT 2014 test set to evaluate the multilingual similarity error rate. This covers English, Czech, French, German, and Spanish, so results between both models are directly comparable. As shown in Table 7, the full model equals or outperforms the one covering the evaluation languages only for all tasks but MLDoc. This suggests that the joint training also yields to overall better representations.

## 6 Conclusions

In this paper, we propose an architecture to learn multilingual fixed-length sentence embeddings for 93 languages. We use a single language-agnostic BiLSTM encoder for all languages, which is trained on publicly available parallel corpora and applied to different downstream tasks without any fine-tuning. Our experiments on cross-lingual natural language inference (XNLI), cross-lingual document classification (MLDoc), and bitext mining (BUCC) confirm the effectiveness of our approach. We also introduce a new test set of multilingual similarity search in 112 languages, and show that our approach is competitive even for low-resource languages. To the best of our knowledge, this is the first successful exploration of general purpose massively multilingual sentence representations.

In the future, we would like to explore alternative encoder architectures like self-attention (Vaswani et al., 2017). We would also like to explore strategies to exploit monolingual data, such as using pre-trained word embeddings, back-translation (Sennrich et al., 2016; Edunov et al., 2018), or other ideas from unsupervised MT (Artetxe et al., 2018c; Lample et al., 2018). Finally, we would like to replace our language-dependant pre-processing with a language-agnostic approach like SentencePiece.[12]

Our implementation, the pre-trained encoder, and the multilingual test set are freely available at `https://github.com/facebookresearch/LASER`.

[12] `https://github.com/google/sentencepiece`

## References

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018c. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, Brussels.

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. *CoRR*, abs/1811.01136.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2017. Weighted set-theoretic alignment of comparable sentences. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 41–45, Vancouver.

Andoni Azpeitia, Thierry Etchegoyhen, and Eva Martínez Garcia. 2018. Extracting parallel sentences from comparable corpora with STACC variants. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*.

Houda Bouamor and Hassan Sajjad. 2018. H2@BUCC18: Parallel sentence extraction from comparable corpora using multilingual sentence embeddings. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*.

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.

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.

Derek F. Wong, Chongman Leong, and Lidia S. Chao. 2018. UM-pAligner: Neural network-based parallel sentence identification model. In *Proceedings of the 11th Workshop on Building and Using Comparable Corpora*.

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.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels.

Andrew M. Dai and Quoc V. Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems 28*, pages 3079–3087.

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, MN.

Philipp Dufter, Mengjie Zhao, Martin Schmitt, Alexander Fraser, and Hinrich Schütze. 2018. Embedding learning through multilingual concept induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1520–1530, Melbourne.

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.

Cristina España-Bonet, Ádám Csaba Varga, Alberto Barrón-Cedeño, and Josef van Genabith. 2017. An empirical analysis of NMT-derived interlingual embeddings and their use in parallel sentence identification. *IEEE Journal of Selected Topics in Signal Processing*, 1340–1348.

Stephan Gouws, Yoshua Bengio, and Greg Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 748–756, Lille.

Francis Grégoire and Philippe Langlais. 2017. BUCC 2017 shared task: A first attempt toward a deep learning framework for identifying parallel sentences in comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 46–50, Vancouver.

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, Belgium.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *CoRR*, abs/1803.05567.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne.

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 *Advances in Neural Information Processing Systems 28*, pages 3294–3302.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based and neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 151–159, Denver, CO.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2535, Copenhagen.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, LA.

Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2017. A survey of cross-lingual embedding models. *CoRR*, abs/1706.04902.

Holger Schwenk. 2018. Filtering and mining parallel data in a joint multilingual space. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–234, Melbourne.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J. Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 2214–2218, Istanbul.

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 30*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels.

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, New Orleans, LA.

Katherine Yu, Haoran Li, and Barlas Oguz. 2018. Multilingual seq2seq training with similarity loss for cross-lingual document classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 175–179, Melbourne.

Zheng Zhang and Pierre Zweigenbaum. 2017. zNLP: Identifying parallel sentences in Chinese-English comparable corpora. In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 51–55, Vancouver.

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.

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 11th Workshop on Building and Using Comparable Corpora*.

## A Training Data

Our training data consists of the combination of the following publicly available parallel corpora:

- **Europarl**: 21 European languages. The size varies from 400k to 2 M sentences depending on the language pair.
- **United Nations**: We use the first 2 million sentences in Arabic, Russian, and Chinese.
- **OpenSubtitles2018:** A parallel corpus of movie subtitles in 57 languages. The corpus size varies from a few thousand sentences to more than 50 million. We keep at most 2 million entries for each language pair.
- **Global Voices:** News stories from the Global Voices Web site (38 languages). This is a rather small corpus with fewer than 100k sentence in most of the languages.
- **Tanzil:** Quran translations in 42 languages, average size of 135k sentences. The style and vocabulary is very different from news texts.
- **Tatoeba:** A community-supported collection of English sentences and translations into more than 300 languages. We use this corpus to extract a separate test set of up to 1,000 sentences (see Appendix C). For languages with more than 1,000 entries, we use the remaining ones for training.

Using all these corpora would provide parallel data for more languages, but we decided to keep 93 languages after discarding several constructed languages with little practical use (Klingon, Kotava, Lojban, Toki Pona, and Volapük). In our preliminary experiments, we observed that the domain of the training data played a key role in the performance of our sentence embeddings. Some tasks (BUCC, MLDoc) tend to perform better when the encoder is trained on long and formal sentences, whereas other tasks (XNLI, Tatoeba) benefit from training on shorter and more informal sentences. So as to obtain a good balance, we used at most 2 million sentences from OpenSubtitles, although more data are available for some languages. The size of the available training data varies largely for the considered languages (see Table 1). This favors high-resource languages when the joint BPE vocabulary is created and the training of the joint encoder. In this work, we did not try to counter this effect by over-sampling low-resource languages.

## B XNLI Results for All Language Combinations

Table 8 reports the accuracies of our system on the XNLI test set when the premises and hypothesis are in a different language. The numbers in the diagonal correspond to the main results reported in Table 2. Our approach obtains strong results when combining different languages. We do not have evidence that distant languages perform considerably worse. Instead, the combined performance seems mostly bounded by the accuracy of the language that performs worst when used alone. For instance, Greek–Russian achieves very similar results to Bulgarian–Russian, two Slavic languages. Similarly, combing French with Chinese, two totally different languages, is only 1.5 points worse than French–Spanish, two very close languages.

## C Tatoeba: Data Set

Tatoeba[13] is an open collection of English sentences and high-quality translations into more than 300 languages. The number of available translations is updated every Saturday. We downloaded the snapshot on November 19, 2018, and performed the following processing: 1) removal of sentences containing ‘‘@’’ or ‘‘http’’, as emails and web addresses are not language-specific; 2) removal of sentences with fewer than three words, as they usually have little semantic information; 3) removal of sentences that appear multiple times, either in the source or the target.

After filtering, we created test sets of up to 1,000 aligned sentences with English. This amount is available for 72 languages. Limiting the number of sentences to 500, we increase the coverage to 86 languages, and 112 languages with 100 parallel sentences. It should be stressed that, in general, the English sentences are not the same for different languages, so error rates are not directly comparable across languages.

## D Tatoeba: Result Analysis

In this section, we provide some analysis on the results given in Table 1. We have 48 languages with an error rate below 10% and 55 with less

[13] `https://tatoeba.org/eng/`

| | | Hypothesis | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | en | ar | bg | de | el | es | fr | hi | ru | sw | th | tr | ur | vi | zh | avg |
| Premise | en | **73.9** | 70.0 | 72.0 | 72.8 | 71.6 | 72.2 | 72.2 | 65.9 | 71.4 | 61.5 | 67.6 | 69.7 | 61.0 | 70.7 | 70.3 | 69.5 |
| | ar | 70.5 | **71.4** | 71.1 | 70.1 | 69.6 | 70.6 | 70.0 | 64.9 | 69.9 | 60.1 | 67.1 | 68.2 | 60.6 | 69.5 | 70.1 | 68.2 |
| | bg | 72.7 | 71.1 | **74.2** | 72.3 | 71.7 | 72.1 | 72.7 | 65.5 | 71.7 | 60.8 | 69.0 | 69.8 | 61.2 | 70.5 | 70.5 | 69.7 |
| | de | 72.0 | 69.6 | 71.8 | **72.6** | 70.9 | 71.7 | 71.5 | 65.2 | 70.8 | 60.5 | 68.1 | 69.1 | 60.5 | 70.0 | 70.7 | 69.0 |
| | el | 73.0 | 70.1 | 72.0 | 72.4 | **72.8** | 71.5 | 71.9 | 65.2 | 71.7 | 61.0 | 68.1 | 69.5 | 61.0 | 70.2 | 70.4 | 69.4 |
| | es | 73.3 | 70.4 | 72.4 | 72.7 | 71.5 | **72.9** | 72.2 | 65.0 | 71.2 | 61.5 | 68.1 | 69.8 | 60.5 | 70.4 | 70.4 | 69.5 |
| | fr | 73.2 | 70.4 | 72.2 | 72.5 | 71.1 | 72.1 | **71.9** | 65.9 | 71.3 | 61.4 | 68.1 | 70.0 | 60.9 | 70.9 | 70.4 | 69.5 |
| | hi | 66.7 | 66.0 | 66.7 | 67.2 | 65.4 | 66.1 | 65.6 | **65.5** | 66.5 | 58.9 | 63.8 | 65.9 | 59.5 | 65.6 | 66.0 | 65.0 |
| | ru | 71.3 | 70.0 | 72.3 | 71.4 | 70.5 | 71.2 | 71.3 | 64.4 | **72.1** | 60.8 | 67.9 | 68.7 | 60.5 | 69.9 | 70.1 | 68.8 |
| | sw | 65.7 | 64.5 | 65.7 | 65.0 | 65.1 | 65.2 | 64.5 | 61.5 | 64.9 | **62.2** | 63.3 | 64.5 | 58.2 | 65.0 | 65.1 | 64.0 |
| | th | 70.5 | 69.2 | 71.4 | 70.1 | 69.6 | 70.2 | 69.6 | 65.2 | 70.2 | 62.1 | **69.2** | 67.7 | 60.9 | 70.0 | 69.6 | 68.4 |
| | tr | 70.6 | 69.1 | 70.4 | 70.3 | 69.6 | 70.6 | 69.8 | 64.0 | 69.1 | 61.3 | 67.3 | **69.7** | 60.6 | 69.8 | 69.0 | 68.1 |
| | ur | 65.5 | 64.8 | 65.3 | 65.9 | 65.3 | 65.7 | 64.8 | 62.1 | 65.3 | 58.2 | 63.2 | 64.1 | **61.0** | 64.3 | 65.0 | 64.0 |
| | vi | 71.7 | 69.7 | 72.2 | 71.1 | 70.7 | 71.3 | 70.5 | 65.4 | 71.0 | 61.3 | 69.0 | 69.3 | 60.6 | **72.0** | 70.3 | 69.1 |
| | zh | 71.6 | 69.9 | 71.7 | 71.1 | 70.1 | 71.2 | 70.8 | 64.1 | 70.9 | 60.5 | 68.6 | 68.9 | 60.3 | 69.8 | **71.4** | 68.7 |
| | avg | 70.8 | 69.1 | 70.8 | 70.5 | 69.7 | 70.3 | 70.0 | 64.7 | 69.8 | 60.8 | 67.2 | 68.3 | 60.5 | 69.2 | 69.3 | **68.1** |

Table 8: XNLI test accuracies for our approach when the premise and hypothesis are in different languages.

| | **ang** | **arq** | **arz** | **ast** | **awa** | **ceb** | **ch** | **csb** | **cy** | **dsb** | **fo** | **fy** | **gd** | **gsw** | **hsb** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **en→xx err.** | 58.96 | 58.62 | 31.24 | 12.60 | 63.20 | 81.67 | 64.23 | 54.55 | 89.74 | 48.64 | 28.24 | 46.24 | 95.66 | 52.99 | 42.44 |
| **xx→en err.** | 65.67 | 62.46 | 31.03 | 14.96 | 64.50 | 87.00 | 77.37 | 58.89 | 93.04 | 55.32 | 28.63 | 50.29 | 96.98 | 58.12 | 48.65 |
| **test sent.** | 134 | 911 | 477 | 127 | 231 | 600 | 137 | 253 | 575 | 479 | 262 | 173 | 829 | 117 | 483 |
| | **jv** | **max** | **mn** | **nn** | **nov** | **orv** | **pam** | **pms** | **swg** | **tk** | **tzl** | **war** | **xh** | **yi** | |
| **en→xx err.** | 73.66 | 48.24 | 89.55 | 13.40 | 33.07 | 68.26 | 93.10 | 50.86 | 50.00 | 75.37 | 54.81 | 84.20 | 90.85 | 93.28 | |
| **xx→en err.** | 80.49 | 50.00 | 94.09 | 10.00 | 35.02 | 75.45 | 95.00 | 49.90 | 58.04 | 83.25 | 55.77 | 88.60 | 92.25 | 95.40 | |
| **test sent.** | 205 | 284 | 440 | 1000 | 257 | 835 | 1000 | 525 | 112 | 203 | 104 | 1000 | 142 | 848 | |

Table 9: Performance on the Tatoeba test set for languages for which we have no training data.

than 20%, respectively (English included). The languages with less than 20% error belong to 20 different families and use 12 different scripts, and include 6 languages for which we have only small amounts of bitexts (less than 400k), namely, Esperanto, Galician, Hindi, Interlingua, Malayam, and Marathi, which presumably benefit from the joint training with other related languages.

Overall, we observe low similarity error rates on the Indo-Aryan languages, namely, Hindi, Bengali, Marathi, and Urdu. The performance on Berber languages (‘‘ber’’ and ‘‘kab’’) is remarkable, although we have fewer than 100k sentences to train them. This is a typical example of languages that are spoken by several millions of people, but for which the amount of written resources is very limited. It is quite unlikely that we would be able to train a good sentence embedding with language specific corpora only, showing the benefit of joint training on many languages.

Only 15 languages have similarity error rates above 50%. Four of them are low-resource languages with their own script and which are alone in their family (Amharic, Armenian, Khmer, and Georgian), making it difficult to benefit from joint training. In any case, it is still remarkable that a language like Khmer performs much better than random with only 625 training examples. There are also several Turkic languages (Kazakh, Tatar, Uighur, and Uzbek) and Celtic languages (Breton and Cornish) with high error rates. We plan to further investigate its cause and possible solutions in the future.

## E Tatoeba: Results for Unseen Languages

We extend our experiments to 29 languages without any training data (see Table 9). Many of them are recognized minority languages spoken in specific regions (e.g., Asturian, Faroese, Frisian, Kashubian, North Moluccan Malay, Piemontese, Swabian, or Sorbian). All share some similarities, at various degrees, with other major languages that we cover, but also differ by their own grammar or specific vocabulary. This enables our encoder to perform reasonably well, even if it did not see these languages during training.