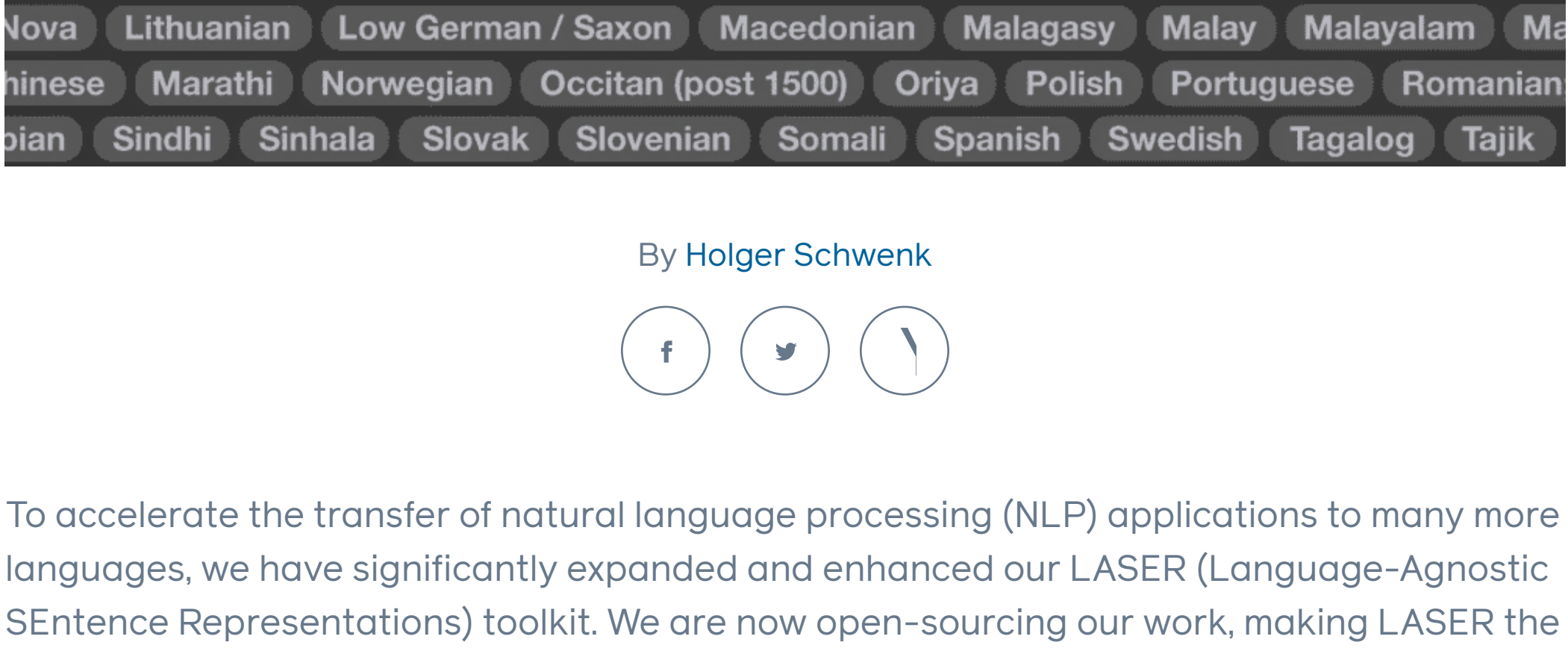
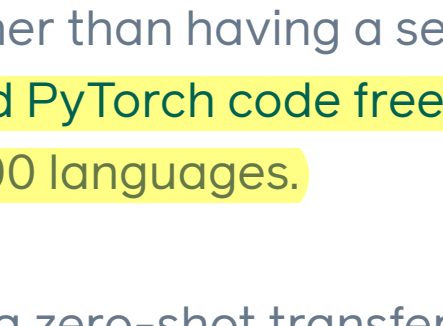


# Zero-shot transfer across 93 languages: Open-sourcing enhanced LASER library



By Holger Schwenk



To accelerate the transfer of natural language processing (NLP) applications to many more languages, we have significantly expanded and enhanced our LASER (Language-Agnostic Sentence Representations) toolkit. We are now open-sourcing our work, making LASER the first successful exploration of massively multilingual sentence representations to be shared publicly with the NLP community. The toolkit now works with more than 90 languages, written in 28 different alphabets. LASER achieves these results by embedding all languages jointly in a single shared space (rather than having a separate model for each). **We are now making the multilingual encoder and PyTorch code freely available, along with a multilingual test set for more than 100 languages.**

LASER opens the door to performing zero-shot transfer of NLP models from one language, such as English, to scores of others — including languages where training data is extremely limited. LASER is the first such library to use one single model to handle this variety of languages, including low-resource languages, like Kabyle and Uighur, as well as dialects such as Wu Chinese. The work could one day help Facebook and others launch a particular NLP feature, such as classifying movie reviews as positive or negative, in one language and then instantly deploy it in more than 100 other languages.

## Performance and feature highlights

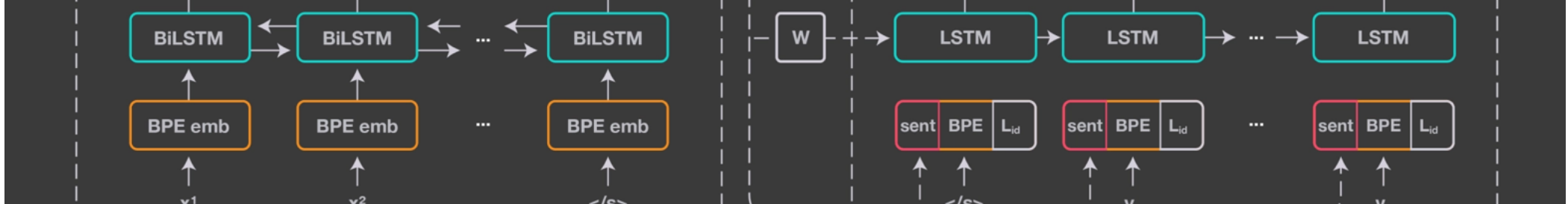
LASER sets a new state of the art on zero-shot cross-lingual natural language inference accuracy for 13 of the 14 languages in the XNLI corpus. It also delivers strong results in cross-lingual document classification (MLDoc corpus). Our sentence embeddings are also strong at parallel corpus mining, establishing a new state of the art in the BUCC shared task for three of its four language pairs. (BUCC is the 2018 Workshop on Building and Using Comparable Corpora.) Along with the LASER toolkit, we are sharing a new test set of aligned sentences in more than 100 languages based on the Tatoeba corpus. Using this data set, our sentence embeddings obtain strong results in multilingual similarity search even for low-resource languages.

LASER also offers several additional benefits:

- It delivers extremely fast performance, processing up to 2,000 sentences per second on GPU.
- The sentence encoder is implemented in PyTorch with minimal external dependencies.
- Languages with limited resources can benefit from joint training over many languages.
- The model supports the use of multiple languages in one sentence.
- Performance improves as new languages are added, as the system learns to recognize characteristics of language families.

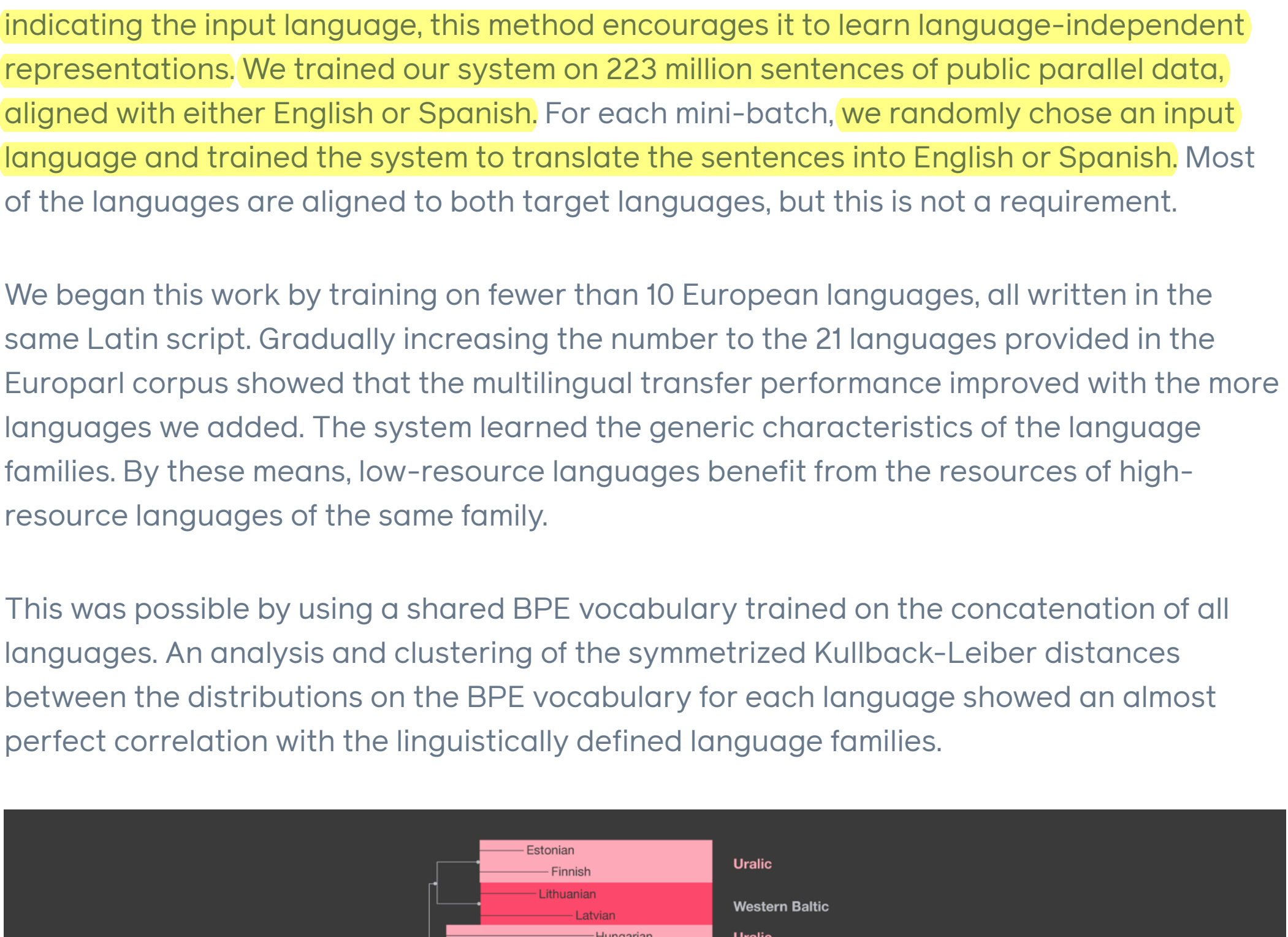
## Universal, language-agnostic sentence embeddings

LASER's vector representations of sentences are generic with respect to both the input language and the NLP task. **The tool maps a sentence in any language to a point in a high-dimensional space with the goal that the same statement in any language will end up in the same neighborhood. This representation could be seen as a universal language in a semantic vector space. We have observed that the distance in that space correlates very well to the semantic closeness of the sentences.**



The image on the left shows a monolingual embedding space. The one on the right illustrates LASER's approach, which embeds all languages in a single, shared space.

Our approach builds on the same underlying technology as neural machine translation: an encoder/decoder approach, also known as sequence-to-sequence processing. **We use one shared encoder for all input languages and a shared decoder to generate the output language.** The encoder is a five-layer bidirectional LSTM (long short-term memory) network. In contrast with neural machine translation, **we do not use an attention mechanism but instead have a 1,024-dimension fixed-size vector to represent the input sentence. It is obtained by max-pooling over the last states of the BiLSTM. This enables us to compare sentence representations and feed them directly into a classifier.**



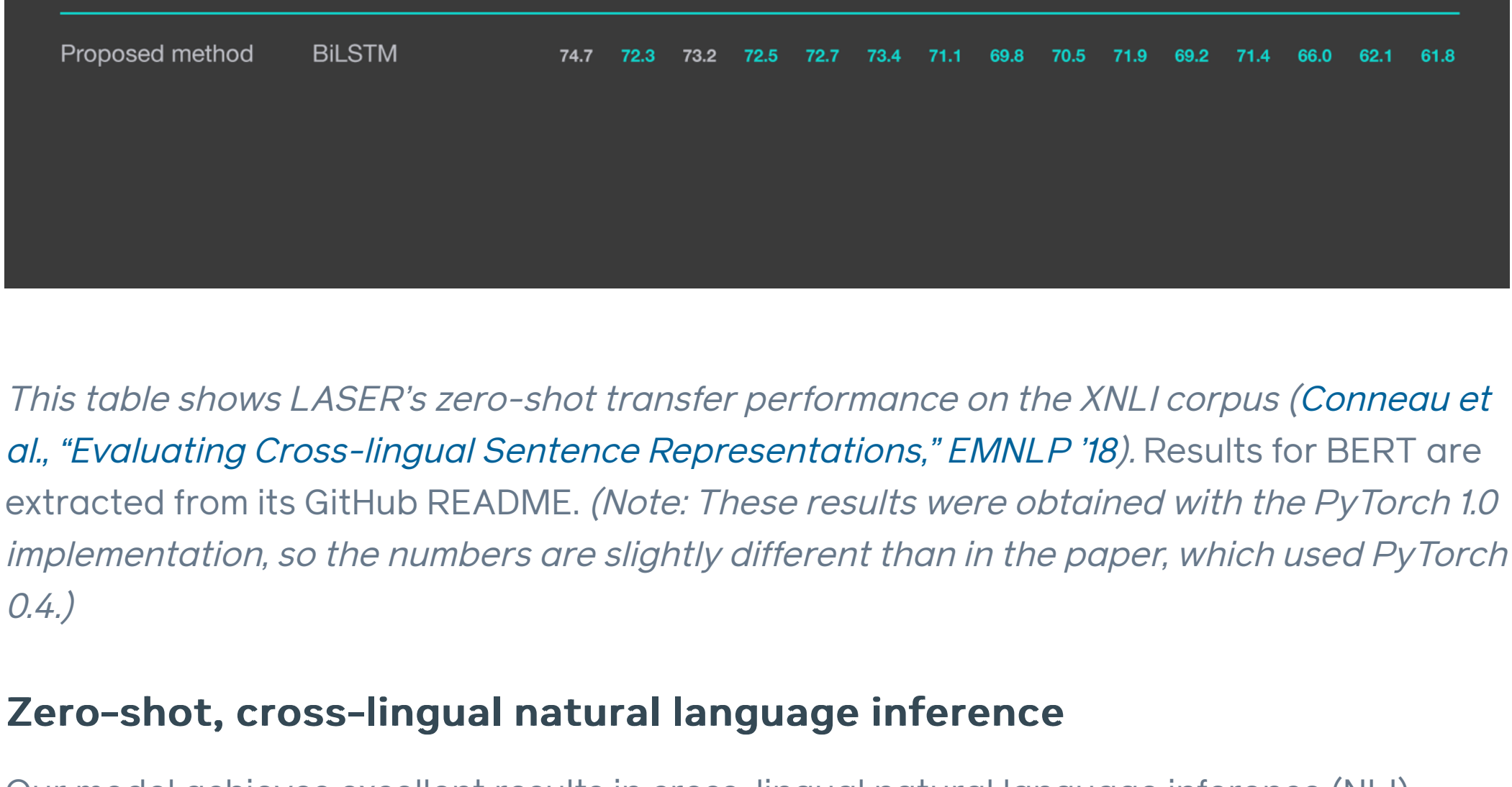
This figure illustrates the architecture of our approach.

**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.** There is no other connection between the encoder and the decoder, as we want all relevant information of the input sequence to be captured by the sentence embedding.

**The decoder must be told which language to generate. It takes a language identity embedding, which is concatenated to the input and sentence embeddings at every time step.** We use a joint byte-pair encoding (BPE) vocabulary with 50,000 operations, trained on the concatenation of all training corpora. **Since the encoder has no explicit signal indicating the input language, this method encourages it to learn language-independent representations. We trained our system on 223 million sentences of public parallel data, aligned with either English or Spanish.** For each mini-batch, **we randomly chose an input language and trained the system to translate the sentences into English or Spanish.** Most of the languages are aligned to both target languages, but this is not a requirement.

We began this work by training on fewer than 10 European languages, all written in the same Latin script. Gradually increasing the number to the 21 languages provided in the Europarl corpus showed that the multilingual transfer performance improved with the more languages we added. The system learned the generic characteristics of the language families. By these means, low-resource languages benefit from the resources of high-resource languages of the same family.

This was possible by using a shared BPE vocabulary trained on the concatenation of all languages. An analysis and clustering of the symmetrized Kullback-Leiber distances between the distributions on the BPE vocabulary for each language showed an almost perfect correlation with the linguistically defined language families.



This graphic illustrates the relationships automatically discovered by LASER between various languages. They correspond very well to the language families manually defined by linguists.

We then realized that a single, shared BiLSTM encoder could handle multiple scripts, and we gradually scaled to all languages for which we identified freely available parallel texts. The 93 languages incorporated into LASER include languages with subject-verb-object (SVO) order (e.g., English), SOV order (e.g., Bengali and Turkic), VSO order (e.g., Tagalog and Berber), and even VOS order (e.g., Malagasy).

Our encoder is able to generalize to languages that were not used (even as monolingual texts) during training. We have observed strong performance on dialects and regional languages, including Asturian, Faroese, Frisian, Kashubian, North Moluccan Malay, Piedmontese, Swabian, and Sorbian. All share some similarities, to various degrees, with other major languages but differ through their own grammar or specific vocabulary.

Thai	Spanish	Neutral
สัปดาห์หน้า, หลานเขาน้องฉันจะมาถือเค้กวันเกิด กันวันเกิดของเขา	Aprender a tocar la guitarra y comenzar una banda era todo lo que hablaba mi sobrino. <i>Learning to play guitar and starting a band was all that my nephew talked about.</i>	(line 4702)
The next week, my nephew asked for an acoustic guitar for his birthday.		

*This table shows how LASER was able to determine the relationship between sentences from the XNLI corpus in different languages. Previous approaches only considered premise and hypothesis in the same language.*

The exact same sentence encoder is also used to mine for parallel data in large collections of monolingual texts. We simply need to calculate the distances between all sentence pairs and select the closest ones. This approach is further improved by considering the margin between the closest sentence and the other nearest neighbors. This search is performed efficiently using Facebook's FAISS library.

We outperform the state of the art on the shared BUCC task by a large margin. The winning system was explicitly developed for this task. We improved the F1 score from 85.5 to 96.2 for

This table shows LASER's zero-shot transfer performance on the XNLI corpus (*Conneau et al., "Evaluating Cross-lingual Sentence Representations," EMNLP '18*). Results for BERT are extracted from its GitHub README. (Note: These results were obtained with the PyTorch 1.0 implementation, so the numbers are slightly different than in the paper, which used PyTorch 0.4.)

## Zero-shot, cross-lingual natural language inference

Our model achieves excellent results in cross-lingual natural language inference (NLI). Performance on this task is a strong indicator of how well the model represents the meaning of a sentence. **We consider the zero-shot setting; in other words, we train the NLI classifier on English and then apply it to all target languages with no fine tuning or target-language resources. For 8 out of 14 languages, the zero-shot performance is within 5 percent of performance on English,** including distant languages like Russian, Chinese, and Vietnamese. We also achieve strong results on low-resource languages like Swahili and Urdu. Finally, LASER outperforms all previous approaches to zero-shot transfer for 13 out of 14 languages.

In contrast to previous methods, which required one sentence to be in English, our system is fully multilingual and supports any combination of premises and hypotheses in different languages.

Premise	Hypothesis	Relation
<b>Bulgarian</b> Никои не знаеше къде отидоха. Their destination was a secret.	<b>Hindi</b> उनका गंतव्य गुप्त था। Nobody knew where they went.	<b>Related</b> (line 210)
<b>Arabic</b> هم، ومن ثم انتقلنا إلى منزل جديد. Um, then we moved to a new house.	<b>Swahili</b> Tulishi kwa nyumba moja maisha yetu yote. We stayed in the same house our whole lives.	<b>Opposite</b> (line 393)
<b>Thai</b> สัปดาห์ที่แล้ว, พี่ชายของฉันถามฉันว่าฉันอยากเล่นกีตาร์อะคูสติกหรือไม่ The next week, my nephew asked for an acoustic guitar for his birthday.	<b>Spanish</b> Aprender a tocar la guitarra y comenzar una banda era todo lo que habléa mi sobrino. Learning to play guitar and starting a band was all that my nephew talked about.	<b>Neutral</b> (line 4702)

This table shows how LASER was able to determine the relationship between sentences from the XNLI corpus in different languages. Previous approaches only considered premise and hypothesis in the same language.

**The exact same sentence encoder is also used to mine for parallel data in large collections of monolingual texts.** We simply need to calculate the distances between all sentence pairs and select the closest ones. This approach is further improved by considering the margin between the closest sentence and the other nearest neighbors. This search is performed efficiently using Facebook's FAISS library.

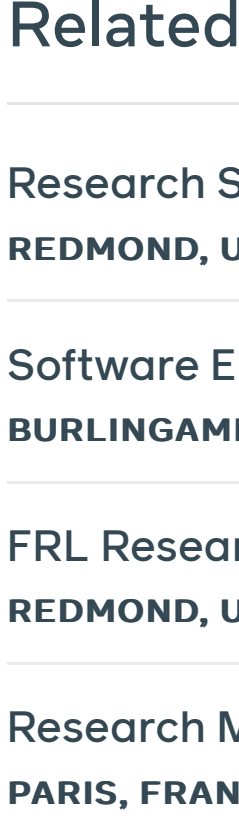
We outperform the state of the art on the shared BUCC task by a large margin. The winning system was explicitly developed for this task. We improved the F1 score from 85.5 to 96.2 for German/English, from 81.5 to 93.9 for French/English, from 81.3 to 93.3 for Russian/English, and from 77.5 to 92.3 for Chinese/English. As these examples show, our results are highly homogeneous across all languages.

(A detailed description of the approach can be found in this research paper co-authored with Mikel Artetxe: **Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond**.)

The same approach can be applied to mine for parallel data in more than 90 languages, using any language pair. This is expected to significantly improve many NLP applications that rely on parallel training data, including neural machine translation in low-resource languages.

## Future applications

The LASER library can also be used for other, related tasks. For example, the properties of the multilingual semantic space can be used for paraphrasing a sentence or searching for sentences with similar meaning — either in the same language or in any of the 93 others now supported by LASER. We will continue to improve our model and add more languages beyond the 93 currently included.

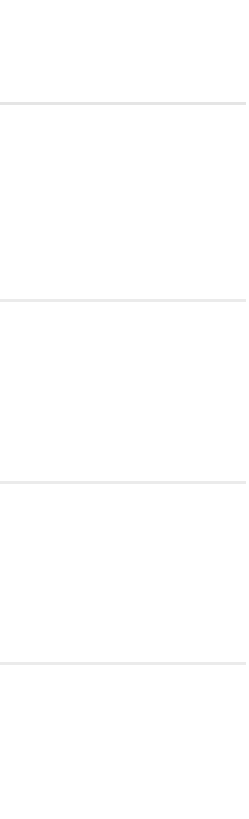


◀ Prev

Facebook open-sources Spectrum 1.0.0 for better mobile image production

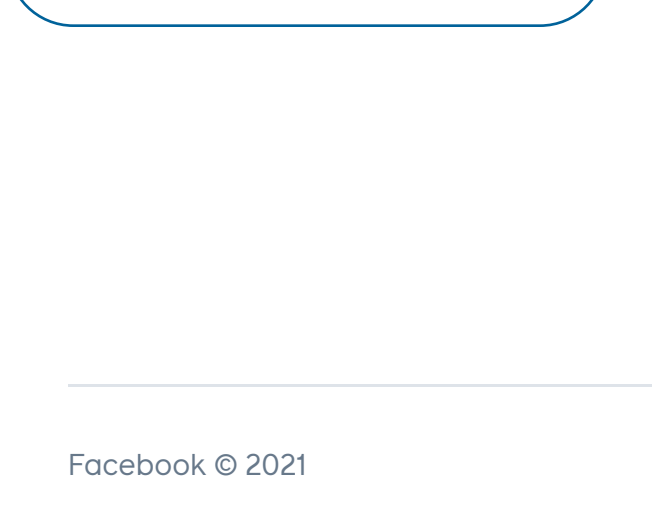
Next ▶

A new predictive model for more accurate electrical grid mapping

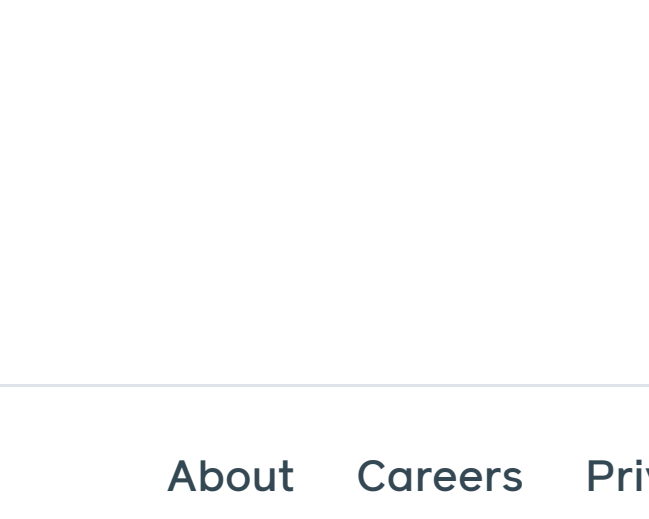


## Read More in AI Research

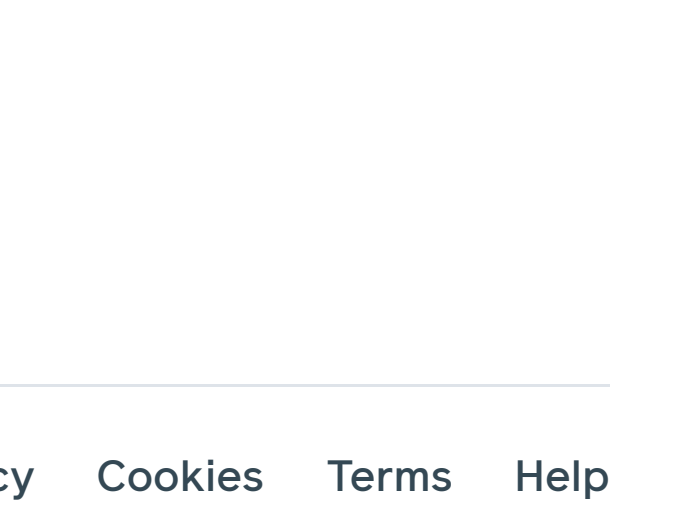
View All ▶



Fully Sharded Data Parallel: faster AI training with fewer GPUs



2019 @Scale Conference recap



Video @Scale 2019 recap



Hydra: A framework that simplifies development of complex applications



Register now for @Scale 2019!



Creating a data set and a challenge for deepfakes

## Related Posts

Research Scientist, Machine Learning for Motion Sensing  
REDMOND, US

Software Engineer, Research Platform  
BURLINGAME, US

FRL Research Software Engineer-Network  
REDMOND, US

Research Manager - Theorem Proving  
PARIS, FRANCE

Research Scientist Manager — Statistical Learning and Experimentation  
TEL AVIV, ISRAEL

[See All Jobs](#)